UCLouvain
Institute for Information
and Communication Technologies,
Electronics and Applied Mathematics

Principled Disentangled Representation Learning: Functional Independence, Identifiability, and Sequential Modeling

Mathieu Cyrille Simon

Thesis submitted in partial fulfillment
of the requirements for the degree of
Ph.D in Engineering Sciences

Dissertation committee:

Prof. Christophe De Vleeschouwer (UCLouvain, Belgium, supervisor)
Prof. Pascal Frossard (EPFL, Switzerland, supervisor)
Prof. Benoît Macq (UCLouvain, Belgium)
Prof. Thomas Maugey (INRIA, France)
Prof. Laurent Jacques (UCLouvain, Belgium, secretary)
Prof. Laurent Francis (UCLouvain, Belgium, chair)

June 15, 2026.

Abstract

Disentangled Representation Learning seeks to recover the latent factors of variation that generate high-dimensional observations, providing a foundation for interpretable, robust, and compositionally aware machine learning. Despite significant empirical progress, the field remains characterized by a gap between heuristic-driven architectures and the formal requirements of identifiability theory. This thesis addresses this gap through a unified framework for principled disentanglement, grounded in functional independence, geometric constraints on the generative mapping, and the inherent structure of sequential data.

A central claim of this thesis is that what makes a latent variable a meaningful factor is fundamentally a geometric property of the generative mapping rather than a statistical property of the data distribution. This perspective provides a new theoretical foundation for disentangled representation learning: it uncovers the structural sources of identifiability, yields formal guarantees that hold without assuming statistical independence between factors, and offers the first principled explanation for why certain model families succeed empirically at disentanglement while others do not.

Sequential data plays a privileged role in this framework. The temporal structure of video provides a natural source of interventional information, and exploiting it carefully reveals structure that static approaches cannot access. In particular, the thesis shows that static and dynamic factors are generically causally dependent, and that respecting this asymmetry, rather than assuming independence, is the key to achieving provable disentanglement in the sequential setting without any form of supervision.

These insights extend naturally beyond strict disentanglement. The same architectural principles that enable the separation of static and dy-

namic content also give rise, when their constraints are relaxed, to compact and expressive global representations of entire video sequences. This connection between disentanglement and representation compactness points toward a unified view of structured generative modeling, in which identifiability and informativeness are two sides of the same design principle.

Keywords: *Representation Learning, Disentanglement, Identifiability, Generative Models, Sequential Data, Unsupervised Learning*

Acknowledgement

A PhD is never the work of one person alone. Behind every equation, every figure, every page of this thesis stands a constellation of people without whom none of this would exist. This is my attempt to thank them, knowing full well that words will fall short.

First and foremost, to my parents, my first and most loyal patrons. You were the ones who made this entire journey possible and gave me the freedom to pursue this path. You have given me your unconditional support, without reservation, every single step of the way. Success is the pride in the eyes of the people we love, and I hope I made you proud. To my sisters, engineering studies are not for the faint-hearted, and there were moments when the demands of this path weighed heavily. You were always there. Your support, your presence, and your ability to make me feel grounded when everything else felt uncertain have meant more than you know.

To Justin and both Quentin, what a strange and wonderful thing it is, to begin a journey surrounded by strangers and to end it surrounded by friends. Sharing this path with you has been one of the unexpected gifts of these years. There is something irreplaceable about people who truly understand what you are going through, not because you explained it, but because they lived it alongside you. To my friends outside the university, the ones who have been there long before any of this began, thank you. You were the ones who kept me anchored to the world beyond research, who showed up, who simply believed in me. Through the late nights, the canceled plans, and the stretches of silence when deadlines swallowed everything, you never wavered. You reminded me, again and again, that there is a life worth living outside of a thesis. I am deeply grateful for each one of you.

To my colleagues, at UCLouvain, thank you to all the members of the

lab who made the daily work richer, more human, and more enjoyable. A research environment is only as good as the people in it, and I was lucky to be surrounded by remarkable ones. During my one-year stay at EPFL's LTS4 lab, I had the privilege of working alongside an exceptional team. I will be honest: I was not always the most present, nor the most radiant presence in the room. The imposter syndrome that shadows so many of us in research was, at times, quite hard to overcome. And yet, every one of you impressed me deeply, the quality of your work, the sharpness of your ideas, the way you carried yourselves through the difficulties of research. I watched it all with a quiet admiration I perhaps never voiced loudly enough. You have all marked this journey, more than you realize.

To Pascaline. No acknowledgement of mine could ever be complete without you, and no words could truly do justice to what you have meant to me throughout this journey. You were there on the hardest days, when doubt was louder than confidence, and on the brightest ones, when everything finally seemed to fall into place. You were my steadiness when I had none, my calm when everything felt overwhelming, my home when the PhD felt like an endless sea. The patience you gave me, the understanding you offered without ever asking for anything in return, the love that never faltered, not once, not even in the moments when I was the most difficult version of myself. I cannot thank you enough for being here. You are, without question, the most important person in these pages.

Finally, to my supervisors, this thesis would not exist without you, and I mean that in the most literal sense. Pascal, thank you for the opportunity to join your lab at EPFL. The brainstorming sessions during those Teams presentations, the detailed feedback, the rigor you brought to every discussion: you pushed me to be better, always. Your standards were high, and somehow you made the pursuit of them feel like a privilege rather than a pressure. I learned an enormous amount from you. Christophe, where do I begin. You have been the steady force on which I could lean when the PhD became overwhelming, and it did, more than once. Your door was always open, quite literally, for long brainstorming sessions at the whiteboard that I will always remember. You are the one who gave me the opportunity to pursue a PhD in the first place, who believed in my ideas when I was not yet sure I believed in them myself, and who supported me every step of the way. Thank you for your availability, your generosity, and your patience.

This thesis represents not just my work, but the culmination of support and guidance from all of you. Thank you.

Author's publication list

- The publications listed below constitute the scientific backbone of this thesis, each directly grounding the key findings presented in the corresponding chapters.

Simon, M. C., Frossard, P., & De Vleeschouwer, C. (2024). Sequential representation learning via static-dynamic conditional disentanglement. In *Proceedings of the European Conference on Computer Vision (ECCV 2024)*.

Simon, M. C., Frossard, P., & De Vleeschouwer, C. (2026). Unsupervised disentanglement without compromises: How functional orthogonality enforces identifiability. In *Proceedings of the Forty-Third International Conference on Machine Learning (ICML 2026)*.

Simon, M. C., Frossard, P., & De Vleeschouwer, C. (2026). All for one and one for all: Learning a global latent representation for video sequences. *Manuscript currently under review*.

- The following publications were produced during the doctoral period but are unrelated to the main contributions of this thesis.

Delinte, N., Gosse, C., Dricot, L., Dessain, Q., Simon, M., Macq, B., & Rensonnet, G. (2021). Microstructural alterations in the white matter of children with dyslexia assessed by multi-fascicle diffusion compartment imaging. In *Proceedings of the International Society for Magnetic Resonance in Medicine (ISMRM 2021)*.

Dessain, Q., Simon, M., Dricot, L., & Macq, B. (2024). Advanced motion correction in diffusion MRI acquisition with gradient cycling. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI 2024)*.

Contents

Contents vii

List of Figures xi

List of Tables xv

List of Acronyms xvii

1 Introduction 1
1.1 Context and Motivation 1
1.2 Research Questions and Contributions 3
1.2.1 Contribution 1: Static-Dynamic Sequential Disentanglement . 4
1.2.2 Contribution 2: Functional Orthogonality and Identifiability . . 5
1.2.3 Contribution 3: Global Latent Representation for Video 6
1.3 Prerequisites and Thesis Outline 8

2 Background 11
2.1 Independent Component Analysis and Source Separation . . . 12
2.2 Disentangled Representation Learning 15
2.3 Causal Modelling for Disentanglement 18
2.4 Evaluating Disentanglement 21
2.5 Generative Architectures for Representation Learning 23
2.5.1 From Autoencoders to Variational Autoencoders and the ELBO 24
2.5.2 Normalizing Flows 32
2.5.3 Hybrid VAE-NF Models 34
2.6 Useful Resources 35

3 Disentangled Sequential Representation Learning 37

3.1 Introduction . . . 38
3.2 Background . . . 40
3.2.1 Disentangled Sequential VAE . . . 40
3.2.2 Other Related Works . . . 42
3.3 Problem Formulation . . . 43
3.4 Method . . . 46
3.4.1 Sufficient Conditions for Disentanglement . . . 46
3.4.2 Practical Implementation . . . 48
3.5 Experiments . . . 51
3.6 Conclusion . . . 55
3.7 Beyond the Framework: Two Emerging Research Directions . . 56
3.7.1 Direction 1 — Multi-Factor Disentanglement on Static Data . 56
3.7.2 Direction 2 — Unconstrained Global Representations for Video 58

4 Unsupervised Disentanglement Without Compromises . . . 61
4.1 Introduction . . . 62
4.2 Related Works . . . 63
4.3 Problem Formulation . . . 64
4.4 Identifiability Conditions . . . 67
4.4.1 Identifiability Under the Independence Assumption . . . 68
4.4.2 Identifiability Under Dependent Factors . . . 69
4.5 Experiments . . . 73
4.6 Discussion and Perspectives . . . 76
4.7 Conclusion . . . 81

5 Learning a Global Latent Representation for Video Sequences 83
5.1 Introduction . . . 84
5.2 Related Works . . . 86
5.3 Global Representation Framework . . . 87
5.4 Method . . . 90
5.5 Experiments . . . 94
5.5.1 Informative global latent s . . . 94
5.5.2 Density estimation . . . 98
5.6 Conclusion . . . 100

6 Conclusion and Perspectives . . . 101
6.1 What This Thesis Has Established . . . 102
6.1.1 A new definition of what a factor is . . . 102

6.1.2 Sequential disentanglement: functional separation without independence . . . 102
6.1.3 Identifiability from functional orthogonality . . . 103
6.1.4 From disentanglement to compact global video representations 104
6.1.5 A unifying thread: theory and practice in dialogue . . . 104
6.2 Limitations and Perspectives . . . 104
6.2.1 Functional Orthogonality: Open Questions at the Frontier . . . 105
6.2.2 Identifiability as a Compass, Not a Destination . . . 106
6.2.3 Global Video Representations: A Promising Direction Still in Its Infancy . . . 107
6.3 Final Thoughts . . . 109

A Supplementary Material for Chapter 3 . . . 111
A.1 Proofs and Detailed Derivations . . . 111
A.1.1 Derivation of the Per-Sequence and Dataset ELBO . . . 111
A.1.2 Proof of Proposition 3.1 . . . 114
A.1.3 Proof of Proposition 3.2 . . . 117
A.1.4 Proof of Proposition 3.3 . . . 119
A.2 Experimental Setup . . . 122
A.2.1 Datasets . . . 122
A.2.2 Disentanglement Metrics . . . 124
A.2.3 Architecture and Hyperparameters . . . 125
A.3 Additional Results . . . 127
A.3.1 Complete Tables of Results . . . 127
A.3.2 Shuffle Ablation and Varying Coefficients . . . 129
A.3.3 Representation Swapping . . . 130
A.3.4 Representation Sampling . . . 134
A.3.5 Interpolation in Latent Space . . . 135
A.3.6 Latent embedding visualization . . . 135
A.4 Details on the Limitations . . . 135

B Supplementary Material for Chapter 4 . . . 143
B.1 Proofs and Detailed Derivations . . . 143
B.1.1 Preliminary Results . . . 143
B.1.2 Proof of Proposition 4.1 . . . 152
B.1.3 Proof of Proposition 4.2 . . . 157
B.1.4 Discussion on Assumption 4.1 and Assumption 4.1 . . . 161
B.2 Disentanglement Under General Latent Domains . . . 163
B.2.1 Proof and Geometric Interpretation . . . 165

B.2.2 The Two-Dimensional Case . . . 170
B.3 Experimental Setup . . . 173
B.3.1 Datasets and Generative Processes . . . 173
B.3.2 Evaluation Metrics . . . 175
B.3.3 Models, Architectures, and Training Objectives . . . 176
B.3.4 Boundary Behavior and Optimization Issues . . . 178
B.3.5 Experiments with Varying Latent Domains . . . 179
B.4 Additional Results and Experiments . . . 180
B.4.1 Quantitative Comparison . . . 180
B.4.2 Qualitative Results . . . 182

C Supplementary Material for Chapter 5 . . . 189
C.1 Dataset Descriptions . . . 189
C.2 Evaluation Metrics and Additional Discussions . . . 191
C.2.1 Label-Based Metrics . . . 191
C.2.2 Label-Free Metrics . . . 193
C.2.3 Density Estimation . . . 196
C.3 Architecture and Hyperparameters . . . 200

Bibliography . . . 207

List of Figures

1.1 Introduction : Overview of the thesis structure. 9

2.1 Background: Illustration of the cocktail party problem. 13
2.2 Background: Illustration of the unidentifiability of nonlinear ICA. 15
2.3 Background: Factors of variations in the 3DShapes dataset [19]. 17
2.4 Background: Comparison of variational autoencoders with different latent structures. 19
2.5 Background: Illustration of the classical AutoEncoder framework and its limitations. 24
2.6 Background: Comparison between AutoEncoder and Variational AutoEncoder. 26
2.7 Background: Complete summary of the standard Variational AutoEncoder framework. 30
2.8 Background: Illustration of the Normalizing Flow models. . . . 33

3.1 Ch.3: Dynamic generation results for the *c-dSprites*, *MPI3D* and *LPCSprites* datasets 39
3.2 Ch.3: Static/dynamic swap results for the MUG dataset. 39
3.3 Ch.3: Schematic view of the proposed model. 47

4.1 Ch.4: Grid Transformations: Visualizing Orthogonal Jacobians. 67
4.2 Ch.4: Football field example for Asm.4.2 70
4.3 Ch.4: MCC and Amari distances quantitative results for $d \in \{3, 6, 9\}$ in the independent/dependent cases. 75
4.4 Ch.4: Qualitative results under various sources domain 76

5.1 Ch.5: Strategies for sequence modeling. 85

5.2 Ch.5: 2D temporal path illustration 88
5.3 Ch.5: Schematic overview of the proposed model. 91
5.4 Ch.5: Samples from the conditional NF decoder. 96

A.1 App.A: Examples of test sequences from *MPI3D*, *c-dSprites* and *LPCSprites*. 124
A.2 App.A: Static/dynamic swap results for the MUG dataset. . . . 132
A.3 App.A: Static/dynamic swap results for the LPCSprites dataset. 132
A.4 App.A: Static/dynamic swap results for the MPI3D dataset. . . 133
A.5 App.A: Static/dynamic swap results for the c-dSprites dataset. 133
A.6 App.A: Static/dynamic swap results for the MHAD dataset. . . 134
A.7 App.A: Static/Dynamic generation results for the LPCSprites dataset. 137
A.8 App.A: Static/Dynamic generation results for the MPI3D dataset. 137
A.9 App.A: Static/Dynamic generation results for the c-dSprites dataset. 138
A.10 App.A: Unconditional generation results for the c-dSprites dataset. 138
A.11 App.A: Interpolation in the latent space of the static code for s-dSprites. 139
A.12 App.A: Interpolation in the latent space of the dynamic code for s-dSprites. 139
A.13 App.A: 2D embedding of the static and dynamic codes computed using t-SNE on all *s-dSprites* test samples, colored according to the different static factors of the dataset. 140
A.14 App.A: 2D embedding of the static and dynamic codes computed using t-SNE on all *s-dSprites* test samples, colored according to the different dynamic factors of the dataset. 140
A.15 App.A: Illustration of the limitations of the two-factor static/-dynamic disentanglement. 141

B.1 App.B: Geometric interpretation of Möbius transformations via stereographic projection. 167
B.2 App.B: Method for constructing counterexamples in 2D. 168
B.3 App.B: Illustration of the condition of Prop.B.1. 169
B.4 App.B: Visualization of a disk automorphism induced by an inversion. 170
B.5 App.B: Qualitative disentanglement results for crescent-shaped domains. 171
B.6 App.B: Example of transformations in 3D. 175
B.7 App.B: Qualitative disentanglement results for independent sources. 183

B.8 App.B: Additional qualitative results for independent sources. . 184
B.9 App.B: Qualitative disentanglement results for dependent sources. 185
B.10 App.B: Additional qualitative results for dependent sources. . . 186
B.11 App.B: Qualitative disentanglement results for non-cubical latent domains. 187
B.12 App.B: Additional results for non-cubical latent domains. 188

List of Tables

3.1 Ch.3: Disentanglement metrics on *s-dSprites*, *MUG* and *MHAD* datasets with independent factors. 53
3.2 Ch.3: Disentanglement metrics on *c-dSprites*, *MPI3D* and *LPC-Sprites* with dependent factors. 54

4.1 Ch.4: Disentanglement performance averaged over 10 random seeds across datasets and models. 80

5.1 Ch.5: Label-based evaluation metrics across datasets. 96
5.2 Ch.5: Label-free evaluation metrics across datasets. 97
5.3 Ch.5: BPD results for the Independent and Autoregressive priors. 99

A.1 App.A: Summary of the factors of variations for the datasets. 123
A.2 App.A: Disentanglement metrics on *s-dSprites*, *c-dSprites*, *MPI3D* and *LPCSprites* 128
A.3 App.A: Disentanglement metrics on *c-dSprites* for varying β and with ablated shuffle operation. 130

B.1 App.B: Averaged disentanglement metrics comparison between NF and VAE. 181

C.1 App.C: Hyperparameters used across all experiments. 203

List of Acronyms

AE Autoencoder (p. 24)

AR Autoregressive (p. 98)

BPD Bits Per Dimension (p. 98)

BSS Blind Source Separation (p. 12)

CIMA Contrastive IMA (regularization loss) (p. 177)

cNF Conditional Normalizing Flow (p. 34)

CNN Convolutional Neural Network (p. 202)

DAG Directed Acyclic Graph (p. 19)

DCI Disentanglement, Completeness, Informativeness (p. 181)

DRL Disentangled Representation Learning (p. 16)

ELBO Evidence Lower BOund (p. 28)

GAN Generative Adversarial Network (p. 42)

GELU Gaussian Error Linear Unit (p. 201)

ICA Independent Component Analysis (p. 12)

ICM Independent Causal Mechanisms (p. 19)

IMA Independent Mechanism Analysis (p. 64)

IS Inception Score (p. 124)

KL Kullback–Leibler (divergence) (p. 27)

KNN K-Nearest Neighbors (p. 191)

LLM Large Language Model (p. 205)

LSTM Long Short-Term Memory (p. 49)

MCC Mean Correlation Coefficient (p. 175)

MI Mutual Information (p. 41)

MIG Mutual Information Gap (p. 181)

MLP Multi-Layer Perceptron (p. 126)

MSE Mean Squared Error (p. 201)

NF Normalizing Flow (p. 32)

NVP Non-Volume Preserving (p. 202)

PCA Principal Component Analysis (p. 69)

PTM Pattern Transformation Manifolds (p. 40)

ReLU Rectified Linear Unit (p. 201)

RMT Riemann Mapping Theorem (p. 170)

SAP Separated Attribute Predictability (p. 181)

SCM Structural Causal Model (p. 18)

STE Straight-Through Estimator (p. 197)

VAE Variational Autoencoder (p. 25)

1 Introduction

1.1 Context and Motivation

A central aspiration of modern machine learning is to build systems that understand the world in the way humans do: not by memorizing high-dimensional signals such as pixel intensities or raw token sequences, but by perceiving the underlying causes that generate those signals. A photograph of a person, for instance, is ultimately the product of a small number of independently controllable, semantically meaningful factors (identity, pose, illumination, expression, and background) combined through the complex, nonlinear optics of image formation. If a learning algorithm could separate and identify these factors automatically, without any supervision, it would gain a degree of interpretability and compositional generalization that purely discriminative models cannot offer. Such a system could transfer knowledge across tasks by manipulating individual factors, generate counterfactual examples, and ultimately reason about causal relationships embedded in data.

This thesis is concerned with *Disentangled Representation Learning* (DRL), the sub-field of unsupervised learning that pursues exactly this goal. A disentangled representation is formally defined as one in which distinct latent variables encode distinct generative factors, so that each variable can be varied independently while leaving the contribution of all other

factors unchanged. Despite this intuitive appeal, the theoretical underpinnings of disentanglement have remained elusive. A landmark negative result by Locatello *et al.* [103] formalized what practitioners had long suspected empirically: without sufficient inductive biases, that is, additional assumptions or constraints, unsupervised disentanglement is *fundamentally impossible*. Infinitely many entangled representations produce the same marginal distribution over observations, rendering the generative model non-identifiable. This impossibility mirrors classical findings in nonlinear Independent Component Analysis (ICA), where recovering independent sources has been proven to fail in the absence of additional structure or supervision [79].

Two broad responses have been explored in the literature. The first introduces *auxiliary information*: partial supervision through class labels, pairs of observations that differ in a known subset of factors, or explicit causal graphical models. While effective, these approaches require annotations or controlled experimental conditions that are rarely available at scale for real-world domains. The second response seeks to identify *structural inductive biases* that are sufficient for identifiability without supervision. The present thesis belongs squarely to this second line of inquiry. Rather than asking what additional data or labels are needed to identify disentangled representations, we ask: *what structural properties of the generative process are sufficient and, more fundamentally, what does it mean for a factor to be a meaningful concept in the first place?*

Three specific gaps motivate the contributions developed here.

Sequential structure as a freely available inductive bias. Most research in DRL has focused on static images. Yet a vast and rapidly growing fraction of real-world data is sequential: video, speech, biomedical time series. Sequential data carries a natural temporal structure that static images do not: the same scene reappears across frames, and the only thing that changes is the dynamic state. This structure constitutes a powerful, freely available source of *interventional information* that, we argue, can be exploited to achieve provable disentanglement without any supervision. Existing methods that address sequential disentanglement have largely treated temporal data through the same statistical lens as static data, failing to formalize and leverage this structure systematically. Moreover, they restrict themselves to the case of *statistically independent* static and dynamic factors: an assumption that is violated in most realistic datasets, as we demonstrate both formally and empirically in this thesis.

A principled, functional notion of what a factor is. The classical formulation of DRL equates disentanglement with the statistical independence of the latent dimensions, drawing direct inspiration from ICA. Yet statistical independence is a fragile proxy for semantic meaningfulness: hair length and gender are intuitively distinct concepts yet statistically correlated in natural datasets; altitude and temperature are independently controllable physical quantities yet they co-vary in nature. Imposing statistical independence on the latent space therefore risks conflating factors that humans consider distinct, or separating factors that are genuinely related through causal or physical mechanisms. A more principled notion of factorhood is needed: one that captures the intuition that factors *act independently on the data* through distinct generative mechanisms, rather than one that demands they be statistically orthogonal in the observed dataset. This functional view was first formalized in [52] and placing it on a general identifiability foundation is one of the central aims of this thesis.

Compact global representations for video. Beyond disentanglement per se, the generative architecture developed to address the sequential setting raises a deeper and largely unexplored question. In that framework, a single shared latent variable summarizes what is common across all frames of a sequence, while separate frame-level codes capture what varies. If the constraint that forces this shared variable to encode only static content is removed, what does the resulting representation encode? The result is still a form of disentanglement, a factorization of the sequence into a single global component and minimal local residuals, but one in which the boundary between the two is no longer drawn at the static/dynamic divide. We establish in this thesis that, under maximum-likelihood training, the global variable naturally absorbs not only static content but also the global structure of the dynamics. This observation opens a third line of inquiry: can a single fixed-size latent variable serve as a compact, general-purpose descriptor of an entire video sequence, enabling downstream tasks without any explicit temporal modelling?

1.2 Research Questions and Contributions

This thesis makes three core contributions in Chap.3–5, each self-contained and adapted from a peer-reviewed publication. In this section, the key research questions and contributions that guided this work are presented.

1.2.1 Contribution 1: Static-Dynamic Sequential Disentanglement (Chap.3)

The first contribution addresses the problem of unsupervised disentanglement of video data into a static code, encoding time-invariant factors, and a set of dynamic codes, encoding temporally varying factors. By leveraging the temporal structure of sequential data as a freely available source of interventional information, we show that provably disentangled representations can be achieved without any auxiliary supervision. This contribution constitutes the backbone of the thesis: it reveals that meaningful factors must be defined through their functional roles in the generative process rather than through statistical independence and constitutes the building block for both subsequent contributions.

Adapted from: M. C. Simon, P. Frossard, and C. De Vleeschouwer. **Sequential representation learning via static-dynamic conditional disentanglement.** *European Conference on Computer Vision (ECCV)*, 2024.

Research questions and contributions.

How can the static and dynamic factors of a video sequence be formally defined, and what causal structure governs their relationship? We propose a new formal definition of static and dynamic factors. The key insight emerging from this formalism is that static and dynamic factors are generically not statistically independent: the static factors may causally influence the possible range and distribution of the dynamic ones. We show that this dependency is present even in simple synthetic settings and is the primary reason existing methods fail on realistic data. By enforcing independence between the two codes, they contradict the inherent causal structure of the data and produce entangled representations.

What generative architecture correctly reflects this causal structure? Rather than extracting the static and dynamic codes independently, as in prior sequential disentanglement work, we introduce a model in which the static code is extracted first via a permutation-invariant aggregated posterior, and the dynamic codes are then obtained, conditioned on the static part. This asymmetric architecture directly encodes the causal relationship between the two factors without requiring prior knowledge of its specific form.

Under what conditions are the ground-truth factors provably identifiable, and can those conditions be enforced without auxiliary supervision? We derive sufficient conditions for the learned codes to identify the ground-truth factors up to invertible reparameterizations, and show that these conditions

are satisfied by a single, architecturally simple operation: randomly permuting the per-frame static code estimates across frames during training. This shuffle constraint penalizes any temporal information that might leak into the static part, while the reconstruction objective simultaneously enforces informativeness. We prove formally that minimizing the resulting model loss is equivalent to satisfying all disentanglement conditions simultaneously, yielding a provably disentangled model without any auxiliary loss, contrastive term, or dataset-specific data augmentation. Experiments on six datasets demonstrate state-of-the-art performance, with particular gains over all competing methods in the practically important setting where static and dynamic factors are statistically dependent.

1.2.2 Contribution 2: Functional Orthogonality and Identifiability (Chap.4)

The second contribution steps back from the sequential setting and addresses a more fundamental question raised by the first: what does it actually mean for a latent variable to correspond to a meaningful factor? In the sequential framework, static and dynamic factors are defined through their functional roles in the generative process, not through statistical independence, and this functional perspective proves essential. The second contribution formalizes and generalizes this insight to the classical, static-data formulation of disentangled representation learning, where it leads to a new theoretical foundation that challenges the prevailing impossibility claims for unsupervised disentanglement. Rather than treating this chapter as a separate line of work, it should be read as the theoretical completion of the first: it provides the principled, general notion of factorhood that the sequential framework implicitly relies upon, and establishes the conditions under which unsupervised identifiability can be achieved.

Adapted from: M. C. Simon, P. Frossard, and C. De Vleeschouwer. **Unsupervised disentanglement without compromises: how functional orthogonality enforces identifiability.** *Forty-Third International Conference on Machine Learning (ICML)*, 2026.

Research questions and contributions.

What is the correct functional definition of a latent concept, and does it lead to an identifiable generative model? We argue that what makes a factor meaningful is not that it is statistically independent of the others, but that it influences the observations through a distinct and separable mechanism. We formalize this as an orthogonality condition on the columns of the Jacobian of the

generative mapping: each latent dimension produces locally independent variations in data space. We prove that under this condition, and when latent factors are statistically independent, the generative model is identifiable up to permutation and element-wise reparameterization. This result substantially extends prior identifiability guarantees to the general case of orthogonal Jacobians, and reveals a direct structural bridge between the nonlinear setting and classical linear ICA.

Can identifiability be achieved when the latent factors are statistically dependent? We show that identifiability can be restored under dependent factors by a complementary global completeness condition: all combinations of factor values are realizable in the dataset. Together, local Jacobian orthogonality and global combinatorial completeness constitute a dual characterization of factorhood without requiring any causal commitment: a factor is a distinct, independently manipulable degree of freedom in the generative process. This dual characterisation does more than restore identifiability: it clarifies what it fundamentally means for a latent variable to constitute a meaningful concept.

Why do Variational Autoencoder (VAE) models succeed at disentanglement empirically, despite the theoretical impossibility result? We provide the first rigorous mechanistic explanation for the empirical success of VAEs at disentanglement. We show that the standard VAE objective implicitly enforces two complementary constraints that together are sufficient for identifiability: one acting globally on the statistical structure of the latent space, and one acting locally on the geometry of the generative mapping. This reframing uncovers a principled source of disentanglement that is broader than independence and more tractable than causality, and provides a unified foundation from which both the empirical success of VAEs and the failure of unconstrained likelihood models can be rigorously explained.

1.2.3 Contribution 3: Global Latent Representation for Video (Chap.5)

The third contribution returns to the sequential setting and explores what happens when the disentanglement objective of the first contribution is deliberately relaxed. In the original framework, a constraint forces the shared latent variable to encode only time-invariant content, producing a clean static/dynamic separation. Removing that constraint does not abandon the underlying generative architecture; it repurposes it. The shared latent variable, no longer confined to static content, is free to absorb whatever

information about the sequence is most useful for reconstruction, including the global structure of the dynamics. This third contribution therefore closes a natural loop: starting from a framework designed for strict disentanglement, it shows that relaxing the disentanglement requirement within the same architecture gives rise to a compact, general-purpose video representation with strong theoretical and empirical properties. It demonstrates that the architectural choices made in service of disentanglement are equally powerful in the unconstrained setting, and arguably more broadly applicable.

Adapted from: M. C. Simon, P. Frossard, and C. De Vleeschouwer. **All for one and one for all: learning a global latent representation for video sequences.** *Manuscript currently under review*, 2026.

Research questions and contributions.

Can a single fixed-size latent variable encode the essential information of an entire video sequence? We propose a generative model in which a single global latent variable $\boldsymbol{s}$, inferred from the full sequence, conditions the reconstruction of each frame independently, without any explicit temporal modelling in the decoder. All information needed to reconstruct the sequence must therefore be concentrated in $\boldsymbol{s}$, with only minimal frame-specific residuals handled locally. Experiments on six diverse video datasets confirm that $\boldsymbol{s}$ captures both the static content and the global dynamic structure of sequences, yielding compact representations that are directly discriminative for downstream tasks without requiring any temporal reasoning at inference time.

Why is the global latent variable encouraged to be maximally informative, and how does this avoid posterior collapse? A theoretical analysis shows that, in contrast to a standard VAE whose objective implicitly penalizes the information content of the latent variable, the proposed model actively encourages $\boldsymbol{s}$ to encode as much of the sequence as possible. This design choice prevents the posterior collapse that plagues high-capacity decoders and provides a theoretical grounding for the informativeness of the learned global representation.

Does conditioning on a global latent improve density estimation, and what are the implications for compression? Conditioning on $\boldsymbol{s}$ consistently improves likelihood across all datasets and temporal settings, confirming that the global latent captures redundancy across frames that would otherwise need to be

modelled repeatedly. This suggests a natural direction toward more efficient video representations, in which shared sequence-level information is encoded once and all frames are decoded relative to this common reference.

1.3 Prerequisites and Thesis Outline

This thesis builds on a core set of theoretical tools that underpin all three contributions.The **Variational AutoEncoders and Normalizing Flows** form the architectural backbone of every model developed here. Their properties, in particular the interplay between reconstruction and regularization in the Evidence Lower Bound, and the invertibility and exact likelihood computation afforded by Normalizing Flows, are exploited throughout and constitute a genuine prerequisite for following the technical developments. **ICA and nonlinear source separation**, together with the associated theory of **identifiability**, provide the conceptual language in which disentanglement is formalized: what it means to recover latent factors, under what conditions this is possible, and where classical guarantees break down. Finally, **causal modeling and inference** recurs across the first two contributions, both as a modeling tool to encode the asymmetric relationship between static and dynamic factors, and as a commitment that is deliberately relaxed in the second contribution to reveal a more general principle. For readers already familiar with these topics, Chap.3 is the natural starting point. For those wishing to consolidate the relevant foundations first, Chap.2 provides a self-contained review before the technical contributions begin.

The logical structure of the thesis and the dependencies between chapters are illustrated in Fig.1.1. Chap.3 constitutes the backbone of the thesis and concludes with a dedicated bridge section that identifies two open questions raised by the sequential framework, thereby motivating the two contributions that follow. Chap.4 addresses the first of these questions, developing a general theory of functional factorhood for static data, while Chap.5 addresses the second, investigating the behavior of the same architecture when the disentanglement constraint is lifted. The two chapters are independent of each other and may be read in either order. Chap.6 synthesizes the contributions of the thesis, discusses their limitations, and outlines the most promising directions for future work. Detailed proofs, ex-

perimental protocols, and additional results are collected in Appendices A, B, and C.

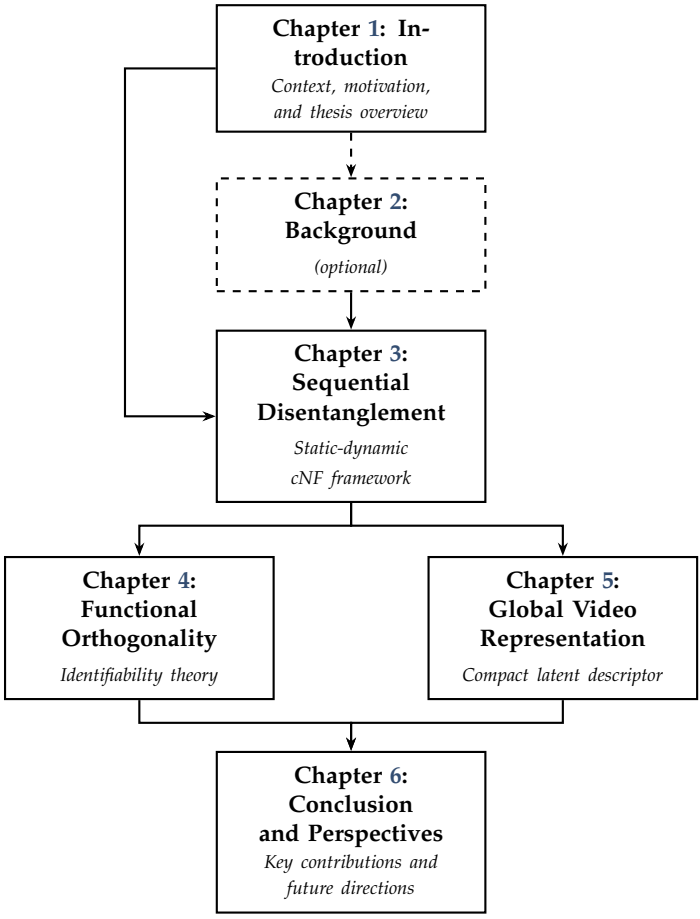


Fig. 1.1 Overview of the thesis structure. The dashed arrow and dashed box indicate that Chapter 2 is optional; readers already familiar with the prerequisites may proceed directly to Chapter 3. Chapter 3 constitutes the backbone from which Chapters 4 and 5 branch independently before converging at the conclusion.

2

Background

Preliminary note:

This chapter lays out the theoretical and methodological foundations on which the contributions of this thesis rest.

Readers already familiar with the following topics may skip directly to Chap.3 or consult only the relevant sections.

- *Independent Component Analysis and Identifiability*
  - *Blind Source Separation and linear ICA (§2.1)*
  - *Nonlinear ICA and the impossibility result (§2.1)*
- *Disentangled Representation Learning*
  - *Formal definition and link to nonlinear ICA (§2.2)*
  - *Landscape of disentanglement notions (§2.2)*
- *Causal Modelling for Disentanglement*
  - *Structural Causal Models and the ICM principle (§2.3)*
- *Generative Architectures*
  - *Autoencoders and Variational Autoencoders (§2.5.1)*
  - *Normalizing Flows (§2.5.2)*
  - *Hybrid VAE-NF models (§2.5.3)*

> We assume the reader has background knowledge in deep learning. Prior literature is not extensively covered here: each core chapter of the thesis contains its own related work discussion that contextualizes our contributions within the existing body of work.

2.1 Independent Component Analysis and Source Separation

Blind Source Separation

Many observations of interest are not elementary signals but rather the result of mixing several underlying *source signals* through an unknown process. *Blind Source Separation* (BSS) is the problem of recovering those sources from the observed mixtures alone, without any knowledge of the mixing mechanism [30].

The canonical illustration is the *cocktail-party problem*, represented in Fig.2.1. Suppose d speakers produce independent speech signals $s_1, \dots, s_d$ simultaneously, and n microphones placed at different positions record the room. At any given time t, each microphone i captures a superposition of all sources:

$$x_i = \sum_{j=1}^{d} a_{ij} s_j, \qquad i = 1, \dots, n, \tag{2.1}$$

or, in matrix form, $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}$, where $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ is the unknown, time-invariant, *mixing matrix* and $\boldsymbol{s} = (s_1, \dots, s_d)^\top$ is the vector of sources at time t. The BSS task is: given only the recordings $\boldsymbol{x}$, recover $\boldsymbol{s}$ without knowing $\boldsymbol{A}$.

Linear ICA

Independent Component Analysis (ICA) [74, 30] solves the BSS problem by exploiting the statistical structure of the sources. The key assumption is that the sources are *mutually statistically independent* and that *at most one* of them follows a Gaussian distribution:

$$p(s_1, \dots, s_d) = \prod_{j=1}^{d} p_j(s_j), \qquad \text{at most one } p_j \text{ Gaussian.} \tag{2.2}$$

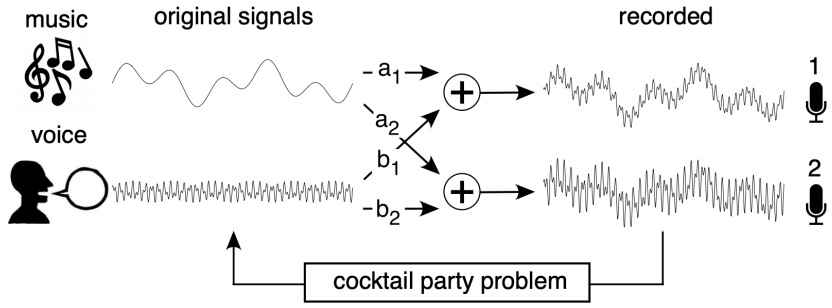


Fig. 2.1 Illustration of the cocktail party problem. The goal is to recover the original sources solely using the microphone recordings. Figure credit - [137]

Under Eq.2.2, one seeks a *demixing matrix* $\boldsymbol{W} \in \mathbb{R}^{d \times n}$ such that $\hat{\boldsymbol{s}} = \boldsymbol{W}\boldsymbol{x}$ has mutually independent components. This is achieved by maximizing a measure of non-Gaussianity or statistical independence of the estimated components [74].

Identifiability. A central concern in any estimation problem is *identifiability*: can the true parameters be recovered uniquely from data? For linear ICA, the answer is affirmative up to two irreducible ambiguities [74]: any valid demixing matrix $\hat{\boldsymbol{W}}$ satisfying Eq.2.2 must satisfy

$$\hat{\boldsymbol{W}} = \boldsymbol{D}\,\boldsymbol{P}\,\boldsymbol{A}^{-1}, \tag{2.3}$$

where $\boldsymbol{P}$ is an arbitrary permutation matrix and $\boldsymbol{D}$ is a non-singular diagonal matrix. In words, the sources can be recovered *up to permutation and scaling*: their order and amplitudes cannot be determined, but each recovered component corresponds to exactly one true source. This ambiguity is intrinsic to the problem, since

$$\boldsymbol{x} = \boldsymbol{A}\,\boldsymbol{s} = \underbrace{\left(\boldsymbol{A}\,\boldsymbol{P}^{-1}\boldsymbol{D}^{-1}\right)}_{\boldsymbol{A}'}\,\underbrace{\left(\boldsymbol{D}\,\boldsymbol{P}\,\boldsymbol{s}\right)}_{\boldsymbol{s}'} = \boldsymbol{A}'\boldsymbol{s}' \tag{2.4}$$

reproduces the observations exactly, with $\boldsymbol{s}' = \boldsymbol{D}\,\boldsymbol{P}\,\boldsymbol{s}$ sharing the same factorized joint distribution as $\boldsymbol{s}$. The two decompositions are therefore statistically indistinguishable. This is nonetheless acceptable, as the recovered signals are merely reordered and rescaled, leaving their content intact and

the disentangled structure of the representation preserved.

The necessity of non-Gaussianity follows directly: if all sources were Gaussian, any rotation of $\hat{\boldsymbol{s}}$ would again be a vector of independent Gaussians, so the demixing matrix would be unidentifiable up to an arbitrary orthogonal transformation, a strictly larger class of ambiguities than Eq.2.3.

Nonlinear ICA

The linear mixing assumption $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}$ is often too restrictive. *Nonlinear ICA* extends the model to an arbitrary smooth, invertible *mixing function*:

$$\boldsymbol{x} = f(\boldsymbol{s}), \qquad f : \mathbb{R}^d \to \mathbb{R}^n, \qquad p(\boldsymbol{s}) = \prod_{j=1}^{d} p_j(s_j). \tag{2.5}$$

The goal is to learn an unmixing function $g : \mathbb{R}^n \to \mathbb{R}^d$ such that $\hat{\boldsymbol{z}} = g(\boldsymbol{x})$ recovers the sources. Identifiability is correspondingly extended: the recovered representation $\hat{\boldsymbol{z}}$ is said to *identify* the sources $\boldsymbol{s}$ if there exist a permutation π and a vector of smooth invertible scalar functions $t_1, \ldots, t_d$ such that

$$\hat{z}_i = t_i\big(s_{\pi(i)}\big), \qquad \forall\, i = 1, \ldots, d. \tag{2.6}$$

This is identifiability *up to permutation and element-wise reparameterization*: each recovered dimension corresponds to exactly one true source, possibly through a monotone scalar transformation. These ambiguities are tolerable since the disentangled structure of the representation is preserved.

The impossibility of nonlinear ICA. Despite this weakened notion of identifiability, the following fundamental result shows that nonlinear ICA is impossible in general without additional structure.

Theorem 2.1 (Adapted from [76]). *Let $\boldsymbol{x} = f(\boldsymbol{s})$ where f is smooth and invertible and $p(\boldsymbol{s}) = \prod_j p_j(s_j)$. Then there exist infinitely many alternative demixing functions $\tilde{g}$ such that $\tilde{\boldsymbol{z}} = \tilde{g}(\boldsymbol{x})$ has independent components, yet $\tilde{\boldsymbol{z}}$ is* not *a permutation or element-wise reparameterisation of the true sources $\boldsymbol{s}$. Consequently, the marginal distribution $p(\boldsymbol{x})$ alone does not identify the true sources.*

Intuitively, any smooth invertible entanglement h of the sources can be absorbed into a redefined mixing function $f' = f \circ h^{-1}$ while leaving the

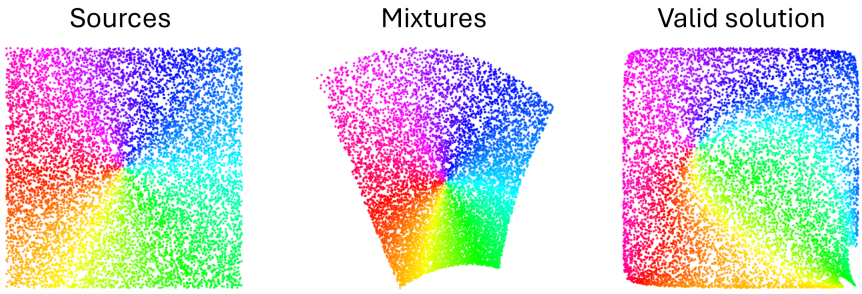


Fig. 2.2 Illustration of the unidentifiability of nonlinear ICA. (Left) Scatterplot of two independent components (Middle) Non linear mixing of the two sources (Right) Valid solution that recover independent components that maps to the observations but do not recover the original sources. Colors are only for reference, to track points across panels.

observed distribution $p(\boldsymbol{x})$ unchanged. The marginal distribution alone therefore contains no information that distinguishes the true independent sources from infinitely many entangled alternatives, rendering the problem fundamentally underdetermined. This is illustrated in Fig.2.2. Identifiability can be restored by conditioning on auxiliary variables $\boldsymbol{u}$ (e.g. segment labels or time indices) and requiring that the conditional source distribution $p(\boldsymbol{s} \mid \boldsymbol{u})$ varies sufficiently across conditions [77, 83].

2.2 Disentangled Representation Learning

Representation Learning

The premise of *representation learning* is that high-dimensional observations $\boldsymbol{x} \in \mathbb{R}^n$ are not uniformly distributed in $\mathbb{R}^n$ but concentrate near a low-dimensional manifold, parameterised by a small number of *factors of variation*, the underlying attributes or causes that govern the appearance of observations. Representation learning aims to map $\boldsymbol{x}$ to a low-dimensional latent vector $\boldsymbol{z} \in \mathbb{R}^d$, $d \ll n$, that makes these factors explicit and accessible for downstream tasks [9].

Disentangled Representation Learning

Disentangled Representation Learning (DRL) [9, 155] strengthens this goal by requiring that each latent dimension z_i encodes a single, semantically distinct factor of variation independently of the others.

Definition 2.1 (Disentangled Representation)**.** Let $\boldsymbol{z}^* = (z_1^*, \ldots, z_d^*)$ be the ground-truth generative factors producing observations $\boldsymbol{x} = f(\boldsymbol{z}^*)$ through a smooth invertible mixing function f. A learned representation $\hat{\boldsymbol{z}} = g(\boldsymbol{x})$ is *disentangled* if it identifies the ground-truth factors in the sense of Eq.2.6: $\hat{z}_i = t_i(z^*_{\pi(i)})$ for all i, for some permutation π and invertible scalar functions $t_1, \ldots, t_d$.

Example: The 3DShapes dataset [19] is a standard DRL benchmark in which images of 3D shapes are generated from six independent ground-truth factors: floor color, wall color, object color, object size, object shape, and camera orientation. An ideal disentangled representation would recover six latent dimensions aligned with these factors, such that varying $\hat{z}_1$ changes only the floor color while all other attributes remain fixed, see Fig.2.3. Such a representation would allow, for instance, to transfer the shape of one object onto the background of another by simply swapping the corresponding latent dimensions, without retraining.

Equivalence with Nonlinear ICA and Impossibility

When the ground-truth factors $z_1^*, \ldots, z_d^*$ are assumed *mutually statistically independent*, the DRL problem becomes formally equivalent to nonlinear ICA: recovering a disentangled representation amounts to inverting the nonlinear mixing function f, which is precisely the task of Eq.2.5. By Theorem 2.1, this is impossible in a fully unsupervised manner. The mirror result for DRL was formalised by Locatello et al. [103], where $f_* p_{\boldsymbol{z}}$ denotes the pushforward of $p_{\boldsymbol{z}}$ through f, i.e. the distribution of $f(\boldsymbol{z})$ when $\boldsymbol{z} \sim p_{\boldsymbol{z}}$.

Theorem 2.2 (Adapted from [103])**.** *For any smooth invertible f and any factorial distribution $p(\boldsymbol{z}^*) = \prod_i p_i(z_i^*)$, and for any disentangled representation g recovering $\boldsymbol{z}^*$ up to the ambiguities of Eq.2.6, there exist infinitely many entangled representations $\tilde{g}$ whose components are* not *element-wise functions of*

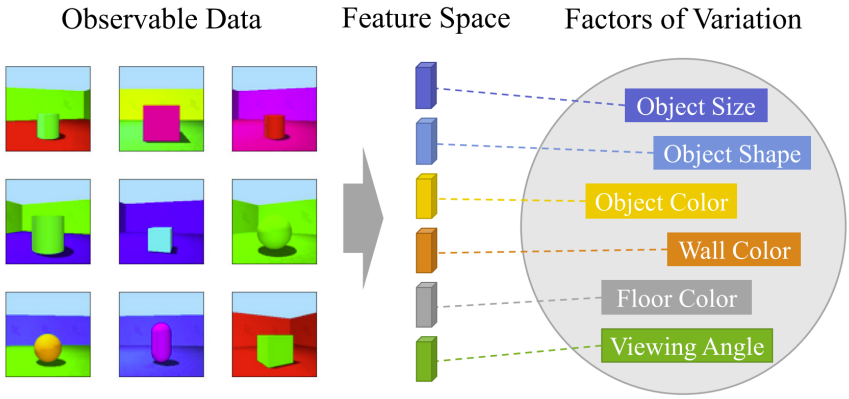


Fig. 2.3 Factors of variations in the 3DShapes dataset [19]. The goal of DRL is to encode these distinct factors with separate latent variables in the feature space. Figure credit - [155]

any permutation of $\mathbf{z}^*$ *such that* $f_* p_{\mathbf{z}^*} = (f \circ \tilde{g}^{-1})_* p_{\tilde{\mathbf{z}}}$. *The observed distribution is identical, leaving the two representations indistinguishable from data alone. Without inductive biases or supervision, no learning algorithm can therefore reliably disentangle the factors of variation.*

Th.2.2 establishes that disentanglement requires additional structure beyond the independence assumption. Similarly to the nonlinear ICA setting, a natural way to recover identifiability is to introduce *supervision* or *auxiliary variables*: providing partial labels, paired observations, or side information that constrains the latent space sufficiently for the factors to be uniquely recoverable [104, 151].

On the notion of disentanglement. It is worth noting that, unlike ICA where the formalization is unambiguous (the "I" stands for *Independent*) there is no single, universally accepted mathematical definition of disentanglement in representation learning. The key intuition is broadly shared: a disentangled representation should *separate the distinct, informative factors of variation* in the data, such that each latent dimension captures one and only one factor. However, different lines of work formalize this intuition differently. The most prevalent approach, reviewed above, equates disentanglement with statistical *independence* of the latent dimensions, drawing

directly from ICA [27, 60, 84, 38, 86, 94, 14, 12]. A separate and mathematically distinct branch grounds disentanglement in *group theory* [59, 179, 21, 126, 169, 154], requiring that the latent space carries a group action that mirrors the symmetries of the generative factors. A third line of work, which we now turn to, moves *beyond independence* by appealing to *causal structure*, arguing that the factors of variation in the data are better characterized by their causal relationships than by their marginal statistics.

2.3 Causal Modelling for Disentanglement

Motivation: Beyond Statistical Independence

The independence assumption underlying the ICA formulation of DRL is not only insufficient for identifiability but also *semantically problematic*. Many semantically distinct concepts are statistically correlated in natural data: hair length and gender co-vary systematically; the speed of an arm in a robot manipulation dataset may depend on the size of the object being held. Forcing statistical independence on the latent space therefore risks either misidentifying the true generative factors or artificially conflating them.

A principled alternative is to model the *causal structure* of the generative process: rather than assuming the factors are independent, one posits that they are connected by causal relationships, and that representation learning should recover both the factors and the causal structure that governs them. A comparison with the standard DRL case is given in Fig.2.4.

Structural Causal Models

A *Structural Causal Model* (SCM) [120, 123, 134] is the standard formal framework for causal modeling. An SCM is a tuple $\mathcal{M} = (\boldsymbol{Z}, \boldsymbol{\epsilon}, \mathcal{F}, p(\boldsymbol{\epsilon}))$ consisting of:

- *Endogenous variables* $\boldsymbol{Z} = (Z_1, \ldots, Z_d)$, representing the latent factors of variation;
- *Exogenous noise variables* $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_d)$, assumed mutually independent: $p(\boldsymbol{\epsilon}) = \prod_i p(\epsilon_i)$;

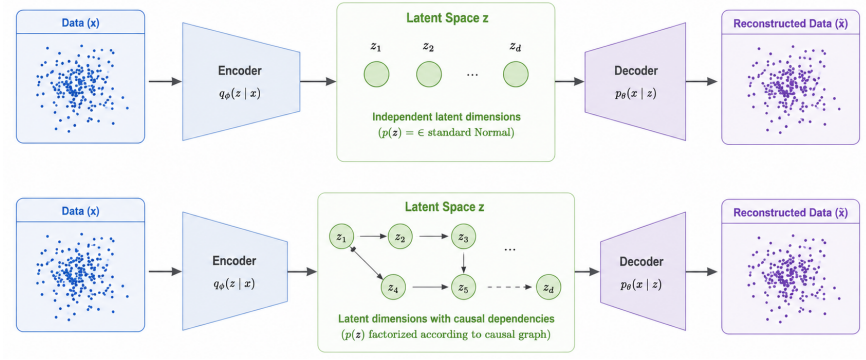


Fig. 2.4 Comparison of Variational AutoEncoders with different latent structures. In the standard VAE (top), latent dimensions are modeled as independent, typically following a factorized prior $p(z) = \mathcal{N}(0, I)$. In the structured variant (bottom), latent variables are organized according to a causal graph, inducing dependencies and a factorized prior aligned with the graph structure. The goal is to recover both the factors and the causal structure that governs them.

- *Structural equations* $\mathcal{F} = \{f_1, \ldots, f_d\}$, one per endogenous variable:

$$Z_i = f_i(\boldsymbol{Z}_{\mathrm{pa}(i)}, \epsilon_i), \tag{2.7}$$

  where $\boldsymbol{Z}_{\mathrm{pa}(i)} \subseteq \boldsymbol{Z}$ denotes the causal parents of Z_i in a directed acyclic graph (DAG) $\mathcal{G}$.

Each equation in Eq.2.7 specifies how Z_i is determined by its causal parents and its own independent noise. All statistical dependencies among the Z_i arise from the DAG structure $\mathcal{G}$; the exogenous noise variables are always mutually independent. When $\mathcal{G}$ has no edges, the factors are independent and the SCM reduces to the standard nonlinear ICA setting. The SCM framework also supports *interventions* $\mathrm{do}(Z_i = c)$, which fix Z_i to a constant value regardless of its parents, enabling counterfactual reasoning about how the system would behave under external manipulations.

The Independent Causal Mechanisms Principle

A central concept in causal modelling, with direct consequences for disentanglement and intervention reasoning, is the *Independent Causal Mechanisms* (ICM) principle [134, 123]. It asserts that the structural functions

$f_1, \ldots, f_d$ in Eq.2.7 are *autonomous modules*: each mechanism f_i is independent of the others in the sense that information about any f_j ($j \neq i$) provides no information about f_i. Modifying one mechanism leaves all others unaffected.

To make this concrete, consider altitude Z_1 causally influencing temperature Z_2 via a mechanism f_2 (the atmospheric lapse rate). The ICM principle asserts that f_2 is an autonomous physical law, independent of the mechanism f_1 governing the distribution of altitudes. Restricting attention to mountainous terrain changes the distribution of Z_1 but not the lapse rate f_2; conversely, altering the atmospheric composition changes f_2 without affecting f_1. This independence is precisely what gives causal models their predictive power under distribution shift and what makes interventions $\mathrm{do}(Z_i = c)$ well-defined: acting on one mechanism leaves all others intact.

The ICM principle defines independence at the level of *generative mechanisms* rather than marginal distributions, and is the causal analogue of the statistical independence assumption in ICA. Applied outside a strict causal graph, it can be formalised as an orthogonality condition on the columns of the Jacobian of the generative function, which is the functional notion of independence developed in Chap.4.

Causal Approaches to DRL and Their Limitations

Building on the SCM and ICM frameworks, several works have proposed to ground disentanglement in causal structure rather than statistical independence, with the aim of recovering both the latent factors and the causal relationships that govern them [167, 16, 100, 165, 150, 95]. While theoretically principled, these approaches carry significant practical burdens: the causal graph $\mathcal{G}$ must either be assumed known or learned jointly with the representation, causal graphs are not generically identifiable from purely observational data without strong additional assumptions [123], and most formal identifiability guarantees require access to *interventional data* (observations collected under known external manipulations) which is rarely available at scale. Furthermore, even when a causal structure exists, it may not align with the factors that are semantically meaningful to a practitioner, and committing to a particular causal graph may obscure rather than clarify the representation. These limitations motivate the search for weaker structural assumptions that do not require a causal commitment yet are still sufficient for identifiable disentanglement, as formalized in Chap.4.

2.4 Evaluating Disentanglement

Having defined what a disentangled representation *is* (Def.2.1), we now address the equally important question of how to *measure* it. Quantifying disentanglement is far from trivial and different intuitions about what makes a representation "disentangled" lead to different, sometimes disagreeing, numerical scores. As a result, the evaluation of disentanglement has itself become an active branch of the literature [20], and no single metric is universally accepted. In practice one reports several complementary metrics that probe different facets of the property.

The classic metrics assume access to a dataset annotated with its ground-truth factors of variation $\boldsymbol{z}^*$, and quantify how well a learned representation $\hat{\boldsymbol{z}} = g(\boldsymbol{x})$ aligns with them. They fall broadly into three families: *alignment-based* metrics that directly compare recovered and true factors (MCC, Amari distance), *information-theoretic* metrics built on mutual information and predictors (MIG, SAP, DCI), and *intervention-based* classifier scores (β-VAE and FactorVAE scores).

Mean Correlation Coefficient (MCC). The MCC [83] is the standard alignment metric in the (nonlinear) ICA and identifiability literature. One computes the pairwise correlation matrix between the d recovered components and the d ground-truth factors, solves a maximum-weight bipartite matching to pair each recovered dimension with a distinct true factor (absorbing the permutation ambiguity), and reports the mean of the matched correlations. A score close to 1 indicates near-perfect recovery.

Amari distance. Originally introduced for linear ICA [6], the Amari distance measures how far the product of the true mixing and the learned unmixing is from a scaled permutation matrix; it vanishes if and only if the two differ by a permutation and a coordinate-wise scaling. A *nonlinear* extension compares the Jacobians of the true generative mapping and the learned inverse, averaged over the data distribution [52]. This Jacobian-based form equals zero exactly when the recovered representation matches the true factors up to the admissible ambiguities of Def.2.1, making it a direct measure of identifiability.

Mutual Information Gap (MIG). The MIG [27] captures the intuition that each ground-truth factor should be encoded by *one and only one* latent dimension. For each factor z_k^*, one computes the mutual information $I(z_k^*; \hat{z}_j)$

with every latent $\hat{z}_j$, and takes the *gap* between the two most informative latents, normalized by the factor entropy:

$$\text{MIG} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{H(z_k^*)}\Big(I(z_k^*;\hat{z}_{j_k^{(1)}}) - I(z_k^*;\hat{z}_{j_k^{(2)}})\Big), \tag{2.8}$$

where $j_k^{(1)}$ and $j_k^{(2)}$ index the highest and second-highest mutual information for factor k. A large gap means the factor is captured almost exclusively by a single dimension. MIG thus specifically rewards *compactness* of the encoding and penalizes a factor being spread across several latents.

Separated Attribute Predictability (SAP). The SAP score [94] follows a similar one-factor-one-dimension philosophy but uses predictability rather than mutual information. For each factor, one builds a score matrix recording how well each individual latent dimension predicts that factor (e.g., via the R^2 of a linear regressor or the accuracy of a simple classifier), and reports the average difference between the two best-predicting dimensions per factor. As with MIG, a high SAP means each factor is predictable from a single dedicated latent rather than from a combination.

DCI: Disentanglement, Completeness, Informativeness. The DCI framework [41] decomposes the quality of a representation into three complementary scores, computed from the importance weights of a regressor (typically a gradient-boosted tree) trained to predict each factor from the full latent code:

- **Disentanglement (D)** — the degree to which each latent dimension is associated with *at most one* factor, measured through the entropy of its importance distribution across factors;
- **Completeness (C)** — the degree to which the information about each factor is concentrated in *at most one* latent, measured through the corresponding entropy across latent dimensions;
- **Informativeness (I)** — the prediction accuracy of the factors from the latent code, measuring how much usable information the representation retains.

DCI is a widely used evaluation framework whose main advantage is that it reports these three aspects separately, rather than summarizing representation quality through a single structural score.

β-VAE score. The β-VAE score [61] is an *intervention-based* metric. One fixes a single factor to a random value and draws *two* batches of observations in which all other factors vary freely; both batches are encoded, and the absolute difference between their mean latent codes is computed and averaged across the batch to form a feature vector. A low-capacity linear classifier is then trained to predict *which* factor was held fixed from this difference vector, and its accuracy is the score. The intuition is that the latent dimension aligned with the fixed factor should exhibit a near-zero difference across the two batches, providing a signal the classifier can exploit; a perfectly disentangled representation makes the fixed factor linearly identifiable from this signal alone.

FactorVAE score. The FactorVAE score [84] refines the β-VAE score to remove a known failure mode (its sensitivity to a poorly chosen classifier and to factors with few values). It uses the *index of the latent dimension with the smallest normalized variance* as the input to a majority-vote classifier predicting the fixed factor. This makes the score more robust and is the variant most commonly reported in recent work [20].

Use in this thesis. The metrics above are employed in Chap.4, where disentanglement is evaluated in the classical static, multi-factor regime: MCC and the Amari distance serve as the primary measures, as they directly probe identifiability in the sense of Def.2.1, with MIG, SAP, DCI, and the β-VAE/FactorVAE scores reported as complementary indicators. Chap.3, by contrast, addresses the *sequential* setting, where disentanglement is coarse-grained into only *two* groups of factors (a time-invariant static code and a dynamic code) rather than many one-dimensional factors, and the object of interest is an entire sequence rather than a single observation. The classical metrics above therefore do not apply directly, and the evaluation framework must be adapted accordingly, following the protocol established in the sequential-disentanglement literature [99, 7, 180, 112]. The corresponding evaluation procedure is detailed in the respective chapter.

2.5 Generative Architectures for Representation Learning

The technical contributions of this thesis rest on two generative frameworks: Variational Autoencoders (VAEs) and Normalizing Flows (NFs).

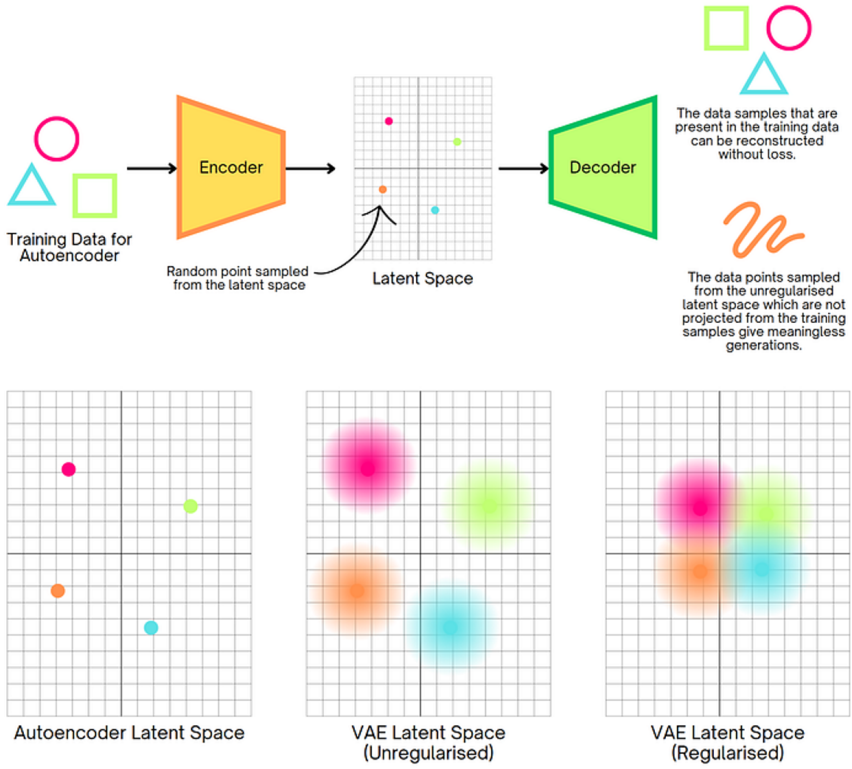


Fig. 2.5 Illustration of the classical AutoEncoder framework and its limitations. Variational Autoencoders address this issue through regularization by introducing a probabilistic framework. Figure credit - Raajan Rajesh Wankhade.

Both provide principled ways to learn a structured latent space from unlabelled data. We introduce them in turn, starting from the simpler deterministic Autoencoder before building up to the full probabilistic picture.

2.5.1 From Autoencoders to Variational Autoencoders and the ELBO

Autoencoders: Learning a Compressed Representation

An *Autoencoder* (AE) [133] is an unsupervised model whose primary goal is to learn a *compressed knowledge representation* of the data. It consists of two parametric functions trained jointly. The *encoder* $\text{Enc}_\phi : \mathcal{X} \to \mathcal{Z}$ maps

an observation $\boldsymbol{x} \in \mathbb{R}^n$ to a low-dimensional latent code $\boldsymbol{z} = \text{Enc}_\phi(\boldsymbol{x}) \in \mathbb{R}^d$ with $d \ll n$, called the *bottleneck representation*. The *decoder* $\text{Dec}_\theta : \mathcal{Z} \to \mathcal{X}$ attempts to reconstruct the original observation from this code alone. Training is driven by a reconstruction loss, typically the mean squared error:

$$\mathcal{L}_{\text{AE}}(\theta, \phi) = \mathbb{E}_{p_{\text{data}}(\boldsymbol{x})} \left\| \boldsymbol{x} - \text{Dec}_\theta\big(\text{Enc}_\phi(\boldsymbol{x})\big) \right\|^2. \tag{2.9}$$

By forcing the data through the bottleneck of dimension d, the encoder is implicitly encouraged to discard redundancy and retain only the most informative structure. AEs are therefore widely used for dimensionality reduction and feature extraction: the latent code $\boldsymbol{z}$ can be thought of as a compact summary of the input, capturing its essential content in a space of much lower dimension.

Fundamental limitations. Despite their utility, AEs are unsuitable as generative models for two related reasons. First, the encoder maps each training point to a *single deterministic code*: there is no distributional structure on the latent space, so a randomly sampled point $\boldsymbol{z}$ will generally not decode to anything meaningful. Second, the latent space is typically *discontinuous*: the decoder is only trained at points that correspond to encodings of training data, and the vast majority of $\mathcal{Z}$ is unconstrained. As a consequence, interpolations between two codes may pass through regions of latent space that the decoder cannot render faithfully, making the AE unsuitable for sampling and generation. This situation is illustrated in Fig.2.5.

These limitations define what one would ideally want from a well-structured latent space:

1. **Continuity.** Similar inputs should have similar latent representations, ensuring smooth transitions: a small perturbation of a code $\boldsymbol{z}$ should yield a small, coherent change in the decoded output.
2. **Completeness.** The model should make use of the entire latent space efficiently, so that any code drawn from $\mathcal{Z}$ decodes to a plausible and diverse output.

The Variational Autoencoder: a Probabilistic Generative Model

The *Variational Autoencoder* (VAE) [88] addresses these limitations by placing a probabilistic framework over the latent space. Rather than encoding an observation as a single point, the encoder outputs the parameters of a distribution over $\mathcal{Z}$, from which a latent code is then sampled (see Fig. 2.6). Just like the AE, the VAE is trained with a reconstruction term, but adds a

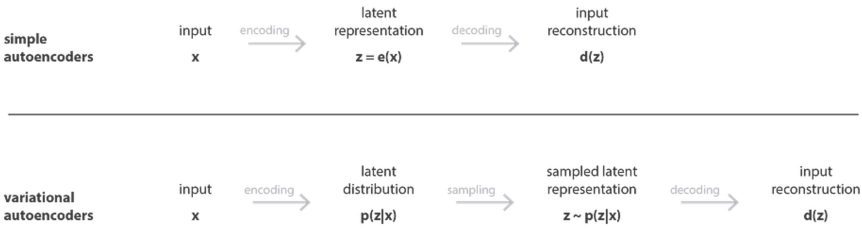


Fig. 2.6 Comparison between AutoEncoder (deterministic) and Variational AutoEncoder (probabilistic). Figure credit - Joseph Rocca.

regularization term that penalizes the learned distribution from deviating too far from a simple Gaussian prior, shaping the latent space into the smooth, well-covered structure the classical AE lacks. It is this regularization that directly enforces the continuity and completeness properties identified above. We now derive the full VAE objective, the *Evidence Lower BOund* (ELBO), and formalize this idea.

Generative model. The VAE defines a joint distribution over observations and latent variables:

$$p_\theta(\boldsymbol{x}, \boldsymbol{z}) = p_\theta(\boldsymbol{x} \mid \boldsymbol{z})\ p(\boldsymbol{z}). \tag{2.10}$$

The two distributions in Eq.2.10 play distinct roles.

- The **prior** $p(\boldsymbol{z})$ encodes our belief about the latent variable *before* observing data. In the standard VAE it is chosen to be an isotropic Gaussian:
$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z};\, \mathbf{0},\, \boldsymbol{I}). \tag{2.11}$$
- The **conditional likelihood** $p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$, also called the *decoder distribution*, describes how an observation $\boldsymbol{x}$ is generated from a latent code $\boldsymbol{z}$. It is parameterized by a neural network with parameters θ.

We observe $\boldsymbol{x}$ but wish to infer the latent variable $\boldsymbol{z}$ that generated it. The quantity of interest is therefore the **posterior distribution** $p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$, which is the encoder's target: given an observation $\boldsymbol{x}$, it tells us which

latent codes z are most likely to have produced it. By Bayes' theorem,

$$p_\theta(z \mid x) = \frac{p_\theta(x \mid z)\; p(z)}{p_\theta(x)}, \tag{2.12}$$

where the denominator is the *marginal likelihood* of the observation:

$$p_\theta(x) = \int p_\theta(x \mid z)\; p(z)\; \mathrm{d}z. \tag{2.13}$$

To train the model we would ideally maximise $\log p_\theta(x)$. However, the integral in Eq.2.13 is *intractable*. The latent space is continuous and potentially high-dimensional: computing $p_\theta(x)$ requires integrating $p_\theta(x \mid z)$ over all possible values of $z \in \mathbb{R}^d$, a problem that grows exponentially harder with d. Furthermore, because the likelihood $p_\theta(x \mid z)$ is parameterised by a nonlinear neural network, the integrand has no closed form. Consequently, neither $p_\theta(x)$ nor the posterior $p_\theta(z \mid x)$ can be computed exactly, and the VAE must resort to *variational inference*.

Variational inference introduces a parametric family of distributions $q_\phi(z \mid x)$ (playing the role of the encoder), called the *approximate posterior*, and optimizes its parameters ϕ so that it is as close as possible to the true (intractable) posterior $p_\theta(z \mid x)$. The quality of this approximation is measured by the Kullback–Leibler (KL) divergence:

$$D_{\mathrm{KL}}\big(q_\phi(z \mid x) \,\|\, p_\theta(z \mid x)\big) = \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{q_\phi(z \mid x)}{p_\theta(z \mid x)}\right] \geq 0, \tag{2.14}$$

with equality if and only if $q_\phi = p_\theta$ almost everywhere.

Deriving the ELBO. We start from the log-likelihood of a single observation x and introduce $q_\phi(z \mid x)$ by multiplying and dividing inside the expectation:

$$\begin{aligned} \log p_\theta(x) &= \log \int p_\theta(x, z)\, \mathrm{d}z \\ &= \log \int q_\phi(z \mid x) \frac{p_\theta(x, z)}{q_\phi(z \mid x)}\, \mathrm{d}z \\ &= \log \mathbb{E}_{q_\phi(z|x)}\left[\frac{p_\theta(x, z)}{q_\phi(z \mid x)}\right]. \end{aligned} \tag{2.15}$$

Applying Jensen's inequality ($\log \mathbb{E}[\cdot] \geq \mathbb{E}[\log \cdot]$ for a concave function)

yields

$$\log p_\theta(\boldsymbol{x}) \;\geq\; \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}\mid\boldsymbol{x})}\right] =: \mathcal{L}_{\text{ELBO}}(\theta,\phi;\boldsymbol{x}). \tag{2.16}$$

The quantity $\mathcal{L}_{\text{ELBO}}$ is called the *Evidence Lower BOund* (ELBO). It is a *lower bound* on the log-likelihood: maximizing it pushes the log-likelihood up while keeping the computation tractable. Expanding the joint distribution using Eq.2.10 gives the standard two-term decomposition:

$$\begin{aligned}\mathcal{L}_{\text{ELBO}}(\theta,\phi;\boldsymbol{x}) &= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{x}\mid\boldsymbol{z})\,p(\boldsymbol{z})}{q_\phi(\boldsymbol{z}\mid\boldsymbol{x})}\right] \\ &= \underbrace{\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}\mid\boldsymbol{z})\right]}_{\text{reconstruction term}} \;-\; \underbrace{D_{\text{KL}}\big(q_\phi(\boldsymbol{z}\mid\boldsymbol{x})\,\|\,p(\boldsymbol{z})\big)}_{\text{regularization term}}.\end{aligned} \tag{2.17}$$

One can verify that the gap between the log-likelihood and the ELBO is exactly the KL divergence between the approximate and true posterior:

$$\log p_\theta(\boldsymbol{x}) = \mathcal{L}_{\text{ELBO}}(\theta,\phi;\boldsymbol{x}) + D_{\text{KL}}\big(q_\phi(\boldsymbol{z}\mid\boldsymbol{x})\,\|\,p_\theta(\boldsymbol{z}\mid\boldsymbol{x})\big). \tag{2.18}$$

Since the KL divergence is non-negative, Eq.2.18 confirms that the ELBO is indeed a lower bound on the log-likelihood. Crucially, it also reveals that maximising the ELBO simultaneously achieves two objectives: (i) it pushes $\log p_\theta(\boldsymbol{x})$ upward, improving the generative model, and (ii) it minimises the KL divergence between the approximate and true posteriors, improving the quality of the encoder. A single tractable objective thus jointly trains both networks.

Interpretation of the two terms.

- The **reconstruction term** $\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}\mid\boldsymbol{z})]$ measures how faithfully the decoder recovers $\boldsymbol{x}$ from a latent code sampled from the approximate posterior. It is the probabilistic counterpart of the deterministic AE reconstruction loss.
- The **regularization term** $D_{\text{KL}}(q_\phi(\boldsymbol{z}\mid\boldsymbol{x})\|p(\boldsymbol{z}))$ penalizes deviations of the approximate posterior from the prior. It prevents the encoder from collapsing to a deterministic point code and forces the aggregate latent distribution to remain close to $\mathcal{N}(\mathbf{0},\boldsymbol{I})$, directly enforcing the *continuity* and *completeness* properties identified as the key failings of the classical AE.

These two objectives are naturally in tension: a more concentrated posterior achieves better reconstruction but deviates further from the prior. The ELBO realises an optimal trade-off between them.

Summary: the VAE in a nutshell.

A VAE is an autoencoder in which both the encoder and the decoder are probability distributions rather than deterministic functions:

- the **encoder** $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ is the *approximate posterior* — a learned distribution over latent codes given an observation;
- the **decoder** $p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$ is the *conditional likelihood* — a learned distribution over observations given a latent code.

The model is trained by maximising the ELBO (Eq.2.17), which serves two complementary roles simultaneously:

1. **Reconstruction.** The reconstruction term forces the encoder decoder pair to recover the input faithfully, just as in a classical AE.
2. **Regularisation.** The KL term forces the approximate posterior $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ to remain close to the prior $p(\boldsymbol{z})$, producing a continuous and complete latent space.

Maximising the ELBO is doubly principled. First, the KL term directly optimises the approximate posterior $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$: by Eq.2.18, maximizing the ELBO minimizes the KL divergence between q_ϕ and the true posterior $p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$, so the encoder genuinely learns to approximate the posterior distribution. Second, since the ELBO is a lower bound on $\log p_\theta(\boldsymbol{x})$, maximising it simultaneously pushes the data log-likelihood upward, making the VAE a *valid generative model* of the data. New observations can therefore be generated by sampling $\boldsymbol{z} \sim p(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and passing the sample through the decoder $p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$. The complete model is summarized in Fig.2.7

Practical Implementation

Parameterization of the approximate posterior. In the standard VAE, the approximate posterior is chosen to be a diagonal Gaussian:

$$q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) = \mathcal{N}\big(\boldsymbol{z};\ \boldsymbol{\mu}_\phi(\boldsymbol{x}),\ \mathrm{diag}(\boldsymbol{\sigma}^2_\phi(\boldsymbol{x}))\big), \tag{2.19}$$

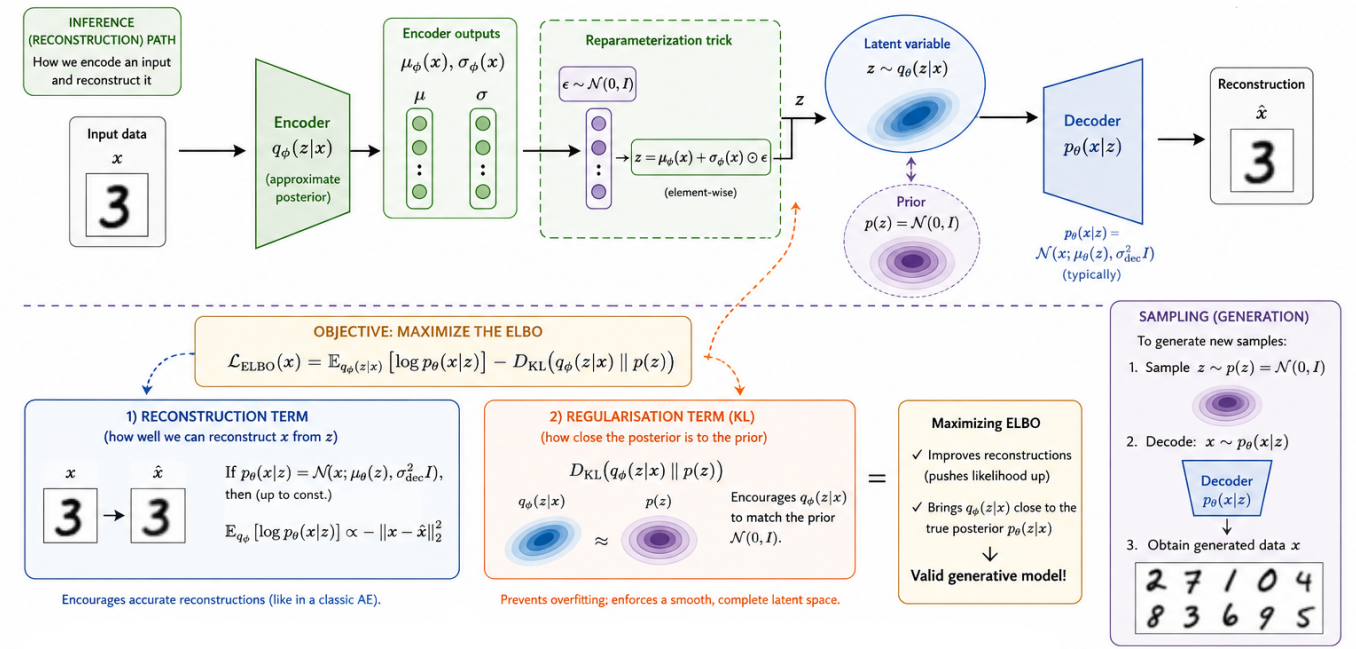


Fig. 2.7 Complete summary of the standard Variational AutoEncoder framework.

where $\boldsymbol{\mu}_\phi$ and σ_ϕ are the outputs of the encoder neural network. Sampling from $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ must be done in a way that allows gradients to flow back through the encoder which is achieved via the *reparameterization trick* [88].

Parameterization of the conditional likelihood. The likelihood $p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$ is also a design choice. The classical selection is a Gaussian with fixed variance:

$$p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) = \mathcal{N}\big(\boldsymbol{x};\, \boldsymbol{\mu}_\theta(\boldsymbol{z}),\, \sigma^2_{\text{dec}} \boldsymbol{I}\big), \tag{2.20}$$

so that $\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) \propto -\|\boldsymbol{x} - \boldsymbol{\mu}_\theta(\boldsymbol{z})\|^2$, recovering an ℓ_2 reconstruction loss. Alternatively, a Bernoulli likelihood yields a binary cross-entropy loss, and a Laplace likelihood yields an ℓ_1 loss, offering flexibility in how reconstruction quality is measured depending on the nature of the data.

VAEs as Natural Disentanglement Models

VAEs occupy a central place in the DRL literature and constitute the backbone of most unsupervised disentanglement methods. Their inductive bias toward disentanglement arises from the *factorized prior*: by encouraging $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ to remain close to $p(\boldsymbol{z}) = \prod_{j=1}^{d} \mathcal{N}(z_j; 0, 1)$, the VAE implicitly promotes statistical independence across latent dimensions.

The β-VAE [61] strengthens this bias by up-weighting the KL term:

$$\mathcal{L}_{\beta\text{-VAE}} = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\big[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})\big] - \beta\, D_{\text{KL}}\big(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p(\boldsymbol{z})\big), \qquad \beta > 1. \tag{2.21}$$

FactorVAE [84] and β-TCVAE [27] further decompose the KL divergence into a sum of the *mutual information* between $\boldsymbol{x}$ and $\boldsymbol{z}$, the *total correlation* across dimensions, and a dimension-wise KL, penalizing the total correlation most heavily to target statistical independence more precisely.

Limitations of VAEs

Despite their widespread use, VAEs suffer from several fundamental limitations. A VAE with a sufficiently powerful decoder can choose to encode all information in the decoder while the encoder collapses to the prior, a regime known as *posterior collapse* or *information preference* [29, 176]: the latent variable $\boldsymbol{z}$ becomes uninformative, the KL term vanishes trivially, and the model degenerates into a pure decoder with no meaningful latent representation. Independently of this, the pixel-wise ℓ_2 reconstruction loss encourages the decoder to produce the statistical mean of all plausible outputs, leading to blurry, low-frequency reconstructions that fall far short of the perceptual quality achievable with modern generative models such as

diffusion models [65].

These limitations interact adversarially with the goal of disentanglement. Strengthening disentanglement by increasing β (Eq.2.21) amplifies the KL penalty, further compressing the latent space and worsening reconstruction quality; relaxing β does the opposite. VAEs therefore face a *fundamental conflict between disentanglement and reconstruction fidelity*, and this tension motivates the use of Normalizing Flows as more expressive decoder distributions throughout this thesis.

2.5.2 Normalizing Flows

A *Normalizing Flow* (NF) [130, 115] is a class of generative model that learns to represent a complex data distribution by transforming a simple base distribution through a sequence of invertible, differentiable mappings. The key idea is that if one knows how to evaluate the density of a simple distribution (e.g. a standard Gaussian) and one applies to it a smooth, invertible function, then the density of the transformed variable can be computed exactly via the *change-of-variables formula*. Unlike VAEs, NFs involve no variational approximation: they compute *exact* likelihoods and are trained purely by maximum likelihood. The model is illustrated in Fig.2.8.

Concretely, let $\boldsymbol{\epsilon} \in \mathbb{R}^d$ follow a simple base distribution $p_{\boldsymbol{\epsilon}}$, typically a standard Gaussian $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and let $f_\theta : \mathbb{R}^d \to \mathbb{R}^d$ be a smooth, invertible mapping parameterized by θ. The transformed variable $\boldsymbol{x} = f_\theta(\boldsymbol{\epsilon})$ then follows the density given by the *change-of-variables formula*:

$$\log p_\theta(\boldsymbol{x}) = \log p_{\boldsymbol{\epsilon}}\big(f_\theta^{-1}(\boldsymbol{x})\big) + \log \left| \det J_{f_\theta^{-1}}(\boldsymbol{x}) \right|, \tag{2.22}$$

where $J_{f_\theta^{-1}}$ is the Jacobian of the inverse transformation. The first term evaluates how likely the corresponding noise $\boldsymbol{\epsilon} = f_\theta^{-1}(\boldsymbol{x})$ is under the base distribution, and the second term $\log |\det J_{f_\theta^{-1}}|$ accounts for the local volume change induced by the mapping: how much f_θ stretches or compresses the space around $\boldsymbol{x}$. This formula is exact, and the model is trained by minimizing the negative log-likelihood:

$$\mathcal{L}_{\text{NF}}(\theta) = -\mathbb{E}_{p_{\text{data}}(\boldsymbol{x})}\big[\log p_\theta(\boldsymbol{x})\big]. \tag{2.23}$$

Only the inverse mapping f_θ^{-1} is needed for training since the loss is evaluated on observed data points $\boldsymbol{x}$. Generation proceeds in the forward di-

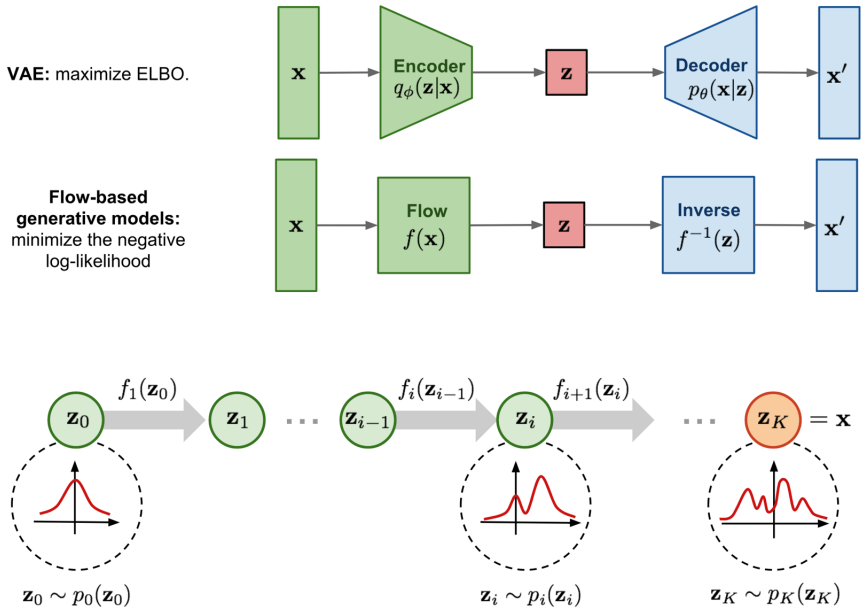


Fig. 2.8 (Top) Comparison of the Normalizing Flow models with the VAE. (Bottom) Illustration of a NF model progressively transforming a simple distribution to a complex one. Figure credit - Lilian Weng.

rection: one samples $\boldsymbol{\epsilon} \sim p_{\boldsymbol{\epsilon}}$ and computes $\boldsymbol{x} = f_\theta(\boldsymbol{\epsilon})$.

The expressiveness of the model depends entirely on the choice of f_θ. Modeling a complex distribution with a single invertible and differentiable function is difficult: the more expressive f_θ becomes, the harder it is to invert and to compute the Jacobian log-determinant. The standard solution is to compose a sequence of simpler invertible transformations each with a tractable Jacobian, $f_\theta = f_K \circ \cdots \circ f_1$. Since a composition of invertible functions is itself invertible, the log-determinant factorises as a sum over layers:

$$\log \left| \det J_{f_\theta^{-1}}(\boldsymbol{x}) \right| = \sum_{k=1}^{K} \log \left| \det J_{f_k^{-1}}\left(\boldsymbol{x}^{(k)}\right) \right|, \tag{2.24}$$

where $\boldsymbol{x}^{(k)}$ denotes the intermediate variable after k inverse steps. By stacking many such layers, the flow can gradually warp the simple base distribution into an arbitrarily complex target distribution.

The key architectural challenge is therefore to design layers that are simultaneously *expressive*, *efficiently invertible*, and have a *tractable log determinant*. A canonical solution is the *coupling layer* introduced in [36], which partitions the input $\boldsymbol{x}$ into two subsets $\boldsymbol{x}_A, \boldsymbol{x}_B$ and applies the transformation

$$\boldsymbol{y}_A = \boldsymbol{x}_A, \qquad \boldsymbol{y}_B = \boldsymbol{x}_B \odot \exp\big(s(\boldsymbol{x}_A)\big) + t(\boldsymbol{x}_A), \tag{2.25}$$

where s and t are arbitrary neural networks. Since $\boldsymbol{x}_A$ passes through unchanged, the inverse is trivially available from $\boldsymbol{y}_A = \boldsymbol{x}_A$, and the Jacobian is lower-triangular so its log-determinant reduces to a simple sum of diagonal entries. Crucially, s and t can be arbitrarily complex networks since they are never inverted. Alternating the partition mask across layers ensures that all dimensions eventually interact. Many other architectures follow this same principle, including NICE [35], Glow [89], autoregressive flows (MAF, IAF, TAR) [116, 90, 175] and Residual Flows [26], each offering different trade-offs between expressiveness, sampling speed, and likelihood evaluation cost.

NFs also extend naturally to *conditional* distributions. A *Conditional Normalizing Flow* (cNF) models $p(\boldsymbol{x} \mid \boldsymbol{c})$ by making the coupling parameters depend on a context variable $\boldsymbol{c}$, replacing $s(\boldsymbol{x}_A)$ and $t(\boldsymbol{x}_A)$ in Eq.2.25 with $s(\boldsymbol{x}_A; \boldsymbol{c})$ and $t(\boldsymbol{x}_A; \boldsymbol{c})$. The change-of-variables formula applies unchanged, preserving exact likelihood evaluation of the conditional distribution.

Limitations: Compared to VAEs, NFs compute *exact* likelihoods, produce *lossless* latent representations (since the encoding f_θ^{-1} is a bijection), and avoid posterior collapse by construction since there is no approximate inference step. However, these advantages come with significant architectural constraints. First, every layer must be invertible *and* have a tractable Jacobian log-determinant, which severely restricts the set of permissible building blocks and rules out many standard deep learning operations. Second, the invertibility constraint forces the input and output to share the same dimensionality, preventing the kind of true dimensionality reduction that an AE or VAE naturally performs.

2.5.3 Hybrid VAE-NF Models

The limitations of VAEs (blurry reconstructions, posterior collapse) and NFs (fixed dimensionality, constrained architectures) are in many ways complementary: each model is strong precisely where the other is weak.

A natural response is to combine them. Crucially, all three probability distributions that appear in the VAE objective in Eq.2.17 (the prior $p(\boldsymbol{z})$, the approximate posterior $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$, and the conditional likelihood $p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$) are themselves probability distributions, and any of them can be replaced by a Normalizing Flow to obtain a strictly more expressive model while preserving the VAE training objective and its latent-variable structure.

Several works have explored each of these three substitutions. Kingma *et al.* [90] and Vahdat *et al.* [147] replace the diagonal Gaussian posterior with a NF-based posterior, capturing richer posterior geometries and tightening the ELBO. Chen *et al.* [29] (Variational Lossy Autoencoder) replace the Gaussian prior with a NF prior; they notably show that a NF prior and a NF posterior are equivalent in terms of training expressiveness, but a NF prior yields a longer sampling path at inference time, empirically producing better samples. Finally, Agrawal *et al.* [2] replace the pixel-wise Gaussian decoder with a conditional NF likelihood, removing the independence assumption across output dimensions and yielding sharper reconstructions.

Role in this thesis. These hybrid architectures improve generative quality but do not constitute disentanglement models in themselves. Throughout this thesis, we draw on all three hybrid strategies (NF posterior, NF prior, cNF likelihood) to build sequential disentanglement models in which the expressivity of Normalizing Flows allows us to simultaneously avoid posterior collapse, obtain sharp reconstructions, and achieve provable disentanglement guarantees.

2.6 Useful Resources

The present chapter provides only a starting point. Each topic covered here is the subject of an extensive literature, on which we have barely scratched the surface. The resources listed below are entry points rather than canonical references; they are selected for accessibility and are particularly suited for readers wishing to deepen their understanding of a specific area before engaging with the technical chapters.

- **Independent Component Analysis.** Hyvärinen *et al.* [76] provide a comprehensive and up-to-date review of nonlinear ICA theory and

algorithms, with particular emphasis on identifiability and self supervised methods. For the classical linear theory, the textbook [74] remains the standard reference.

- **Representation Learning and Disentanglement.** Wang *et al.* [155] offer a systematic survey of disentangled representation learning, covering independence-based, causal, and group-theoretic approaches with a unified perspective.

- **Causal Representation Learning.** Schölkopf *et al.* [134] set out the program for grounding representation learning in causal principles, and provide an accessible entry into Structural Causal Models and the Independent Causal Mechanisms principle from a machine learning perspective.

- **Variational Autoencoders.** For intuitive, implementation-oriented introductions to VAEs and the ELBO, two complementary tutorials are available: Raajan Rajesh Wankhade (link) focuses on implementation, while Joseph Rocca (link) provides a deeper conceptual and geometric perspective. Both are helpful complements to the more formal treatment given here.

- **Normalizing Flows.** Two accessible introductions are available: the UvA Deep Learning tutorial by Phillip Lippe (link) provides a hands-on approach with worked examples of coupling layers and image modelling, and the blog post by Lilian Weng (link) offers a concise and well-illustrated survey of flow architectures. For a thorough theoretical treatment, the survey by Papamakarios *et al.* [115] covers the full landscape of flow architectures and their properties.

3

Sequential Representation Learning via Static-Dynamic Conditional Disentanglement

Adapted from:

Simon, M. C., Frossard, P., & De Vleeschouwer, C. (2024). Sequential representation learning via static-dynamic conditional disentanglement. In *Proceedings of the European Conference on Computer Vision (ECCV 2024)*.

3.1 Introduction

Disentangled Representation Learning (DRL) focuses on embedding high dimensional complex data into a low dimensional space which factorizes the hidden underlying factors of variations. This disentanglement is expected to facilitate numerous downstream generation or classification tasks and enhance model explainability [105, 102, 156, 44, 9]. The present work in particular is concerned with unsupervised sequential data DRL [149, 143, 146] which, compared to disentanglement on static data [88, 156, 179, 104, 14, 110, 60], exhibits specific temporal structure that could be leveraged in the learning process. More precisely, the objective is to learn a representation of video data that factorizes the time independent factors (i.e., what is *static*; the identities, the backgrounds) and time varying factors (i.e., what is *dynamic*; the poses, the motions) in two separate feature vectors. In Fig.3.2, we show how this disentanglement allows for example to swap the facial expressions (dynamic factors) between faces (static factors).

As the above description might leave room to some ambiguity into what constitutes a static/dynamic variable, we introduce in this work an explicit and novel formal definition of those factors. Through this formalism, we notably unveil that the factors might not be independent. This work proposes to investigate the effect of that dependency on the disentanglement models and shows that it translates in difficulties for SotA methods to accurately capture a representation of static and dynamic features. Their extracted codes remain entangled, highligting the importance of developing a robust disentanglement framework that also generalizes to non independent factors. To solve those issues, we propose a new generative model that explicitly incorporates the causal relationship between static and dynamic factors. Diverging from the classical Variational AutoEncoder (VAE) framework, it leverages Conditional Normalizing Flow (cNF) [2, 161, 111] to offer improved disentanglement and model expressivity. By exploiting our new formalism, we uncover conditions under which the ground truth factors can be provably inferred and show that our model directly satisfies those conditions using a simple shuffle operation. This leads to a novel evidence lower bound that naturally encourages disentanglement and makes the learned representation provably disentangled.

Among the other works that have addressed this objective under the VAE framework [34, 69, 10, 4], DSVAE [99] is one of the first models that

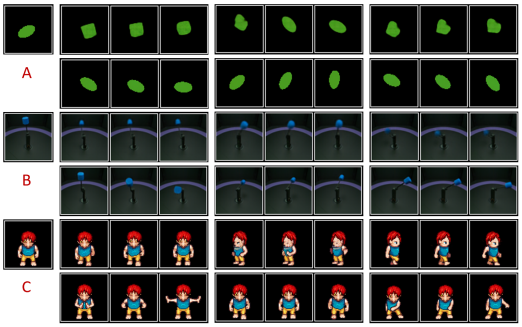


Fig. 3.1 Dynamic generation results for the *c-dSprites* (A), *MPI3D* (B) and *LPCSprites* (C) datasets. The sequences are generated by fixing the static code to the value given by the left image and sampling the dynamic variables from the prior. For each dataset, the first row corresponds to samples from the competing CDSVAE model [7] and the second row to samples from our proposed model.

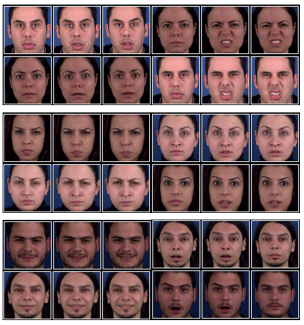


Fig. 3.2 Static/dynamic swap results for the MUG dataset. Odd rows are test input sequences while even rows are sequences generated by swapping the static and dynamic codes of the test sequences using our model.

proposed unsupervised disentanglement of the static and dynamic factors through a factorized sequential latent variable model. However, it was proven to fail in properly disentangling the factors, and to result in representations that heavily depend on the architecture and hyperparameters [106]. Haga *et al.* [54] propose to alleviate this problem using action labels as supervision. However, this requires costly annotations that are rarely available. Subsequent works have therefore extended the DSVAE framework using self-supervised methods aimed at constraining the disentanglement [7, 112, 180, 56, 106, 144]. Although these approaches demonstrate improvements over DSVAE, they rely on increasingly complex and cumbersome constraints, besides limiting themselves to cases with independent static/dynamic factors. Comparatively, our approach solves both these problems with our shuffle constraint that does not require any additional loss hence making our method highly simple and non dataset/domain specific. The model is validated on both synthetic and real datasets showing state-of-the-art results on multiple disentanglement metrics.

3.2 Background

Let $\mathcal{G} = \{\boldsymbol{x}^j_{1:T_j}\}_{j=1}^N$ denote a dataset of N video sequences of length T_j where each $\boldsymbol{x}^j_t$ corresponds to a video frame at time step t in sequence j. The main hypothesis underlying this work is that the frames $\boldsymbol{x}_t$ can be uniquely represented into two parts, $\boldsymbol{s}$ and $\boldsymbol{d}_t$ with:

- $\boldsymbol{s}$: the *static* factors, common and shared by all frames, representing the *time invariant* information;
- $\boldsymbol{d}_t$: the *dynamic* factors, changing between frames, corresponding to the *time varying* information.

Due to temporal coherence, consecutive frames have common factors that are shared between them. Therefore, inspired by Pattern Transformation Manifolds (PTM) representation [152], we consider that each frame resides on a common manifold that corresponds to that shared content. The code $\boldsymbol{s}$ represents the manifold while $\boldsymbol{d}_{1:T}$ captures the varying parts specific to each frame. For example, for a video of a person walking, all frames share the same background and person identity thus $\boldsymbol{s}$ represents those static factors, while $\boldsymbol{d}_{1:T}$ encodes the person specific pose at each frame i.e., the motion.

3.2.1 Disentangled Sequential VAE

With the above representation, Li and Mandt [99] proposed the following probabilistic model where each frame is generated from its corresponding static and dynamic codes and the dynamic codes are obtained given previous time steps:

$$p(\boldsymbol{x}_{1:T}, \boldsymbol{s}, \boldsymbol{d}_{1:T}) = p(\boldsymbol{s}) \prod_{t=1}^{T} p(\boldsymbol{d}_t|\boldsymbol{d}_{<t}) p(\boldsymbol{x}_t|\boldsymbol{s}, \boldsymbol{d}_t). \tag{3.1}$$

Reading the factors in order mirrors the generative process: the static code $\boldsymbol{s}$ is drawn once for the whole sequence, each dynamic code $\boldsymbol{d}_t$ is generated from its past $\boldsymbol{d}_{<t}$ to enforce temporal coherence, and each frame $\boldsymbol{x}_t$ is emitted from its static and dynamic codes. The objective is to extract the codes $\boldsymbol{s}$ and $\boldsymbol{d}_{1:T}$ given the observed data i.e., to learn the factorized posterior distribution $p(\boldsymbol{s}, \boldsymbol{d}_{1:T}|\boldsymbol{x}_{1:T})$. Variational inference can be used to learn an approximation of that posterior:

$$q(\boldsymbol{s}, \boldsymbol{d}_{1:T}|\boldsymbol{x}_{1:T}) = q(\boldsymbol{s}|\boldsymbol{x}_{1:T}) \prod_{t=1}^{T} q(\boldsymbol{d}_t|\boldsymbol{s}, \boldsymbol{d}_{<t}, \boldsymbol{x}_{\leq t}) \tag{3.2}$$

with the two terms obtained through separate sequential networks [99]. The model is trained via the VAE algorithm. This encoder network, introduced in [99], is named 'full q' and is the building block of all following sequential disentanglement works. Unfortunately, despite the disentanglement bias imposed by the factorized latent representation, this model might not be sufficient to achieve a well-disentangled representation. Indeed, as formalized in [103], unsupervised disentanglement is fundamentally impossible to achieve without inductive biases since there exists a potentially infinite number of entangled $\boldsymbol{s}$ and $\boldsymbol{d}_{1:T}$ having the same marginal $\boldsymbol{x}_{1:T}$, making the generative model non identifiable. Additionally, VAEs are known to suffer from the so-called 'information preference property' [29, 176]. A model with a complex conditional likelihood distribution tends to model a maximum of information in the decoder and ignores the latent variables that become non-informative and trivially match the prior.

DSVAE [99] tries to mitigate that problem using low dimensional dynamic codes $\boldsymbol{d}_{1:T}$ to promote information in the static variable. However, this requires careful adaptation of the hyperparameters making this method highly impractical without any guarantees on the resulting representation. Following works have therefore extended the DSVAE framework in order to constrain the disentanglement using self-supervised methods that augment the model loss with mutual information (MI) terms $I(\boldsymbol{u}, \boldsymbol{v}) = D_{\mathrm{KL}}\big(p(\boldsymbol{u}, \boldsymbol{v}) \,\|\, p(\boldsymbol{u})p(\boldsymbol{v})\big) \geq 0$, which measure the statistical dependence between two variables, vanishing iff they are independent. More specifically, the aim is to maximize $I(\boldsymbol{x}_{1:T}, \boldsymbol{s})$ and $I(\boldsymbol{x}_{1:T}, \boldsymbol{d}_{1:T})$ while simultaneously minimizing $I(\boldsymbol{d}_{1:T}, \boldsymbol{s})$. Estimating those MI terms is not straightforward and is done either through domain specific methods (e.g., data augmentation, adversarial training or external models [7, 180, 56]), or through the model itself (e.g., costly encoding-decoding counterfactual losses [144] or contrastive MI estimation with VAE-based sampling [112]). Additionally, noting an apparent improvement on disentanglement, these methods modify Eq.(3.2) in order to model both codes independently by ablating the conditional link.

The above models however have several limitations:

1. While the proposed constraints have shown to improve over DSVAE, these methods are either modality-based i.e., data/task-dependent, or they rely on the inherent ability of their model architecture to naturally disentangle. Thus, might end up promoting an improperly disentangled representation.

2. Even for simple datasets, semantically meaningful factors are generally not independent [167]. Hence, the conditional link between the two codes should *not* be ablated. We argue that the improvement observed from ablation is due to a lack of expressivity in the conditioning model and MI constraints that are contradicting under potential conditionality.
3. The definition of the static and dynamic factors is not formal, leaving room to some ambiguity into what constitutes them.

3.2.2 Other Related Works

Differently than the VAE-based methods derived from the DSVAE model [99, 112, 7, 56, 180, 144, 4, 106, 54], Berman *et al.* [10] recently developed structured Koopman autoencoders that try to achieve multifactor disentanglement of the sequential data with more than two disentangled components. However, their results still lie behind recent DSVAE-based models [112]. In parallel, the work FHVAE [69, 68] also proposed unsupervised disentanglement through an LSTM-based model but is limited to audio data. Then, within the works that try to tackle disentanglement learning for static data [27, 84, 60], we can mention the field of style/content disentanglement [14, 151, 46, 173], with a definition of the static variables related to ours (Sec.3.3). Interestingly, the effectiveness of those methods can thus be explained through the prism of our formalism in Sec.3.4.1. This further demonstrates our propositions and justifies our approach for sequential disentanglement, which comparatively does not require additional losses [151, 46] or model settings that depend on the dataset [173]. Finally, several GAN-based models have been successfully applied for disentangling style/content [81, 82] or appearance/motion for video generation [146, 143, 157]. However, compared to VAE approaches, these methods do not allow to easily encode the input frames and are thus typically less applicable in the context of representation learning.

For image generation, several works have also explored improving VAE models expressivity using NFs to augment either the approximated posterior [90, 147], the prior [29] or the conditional likelihood [161, 2]. For videos, Marino *et al.*[108] tried to augment the conditional likelihood of sequential latent variable models based on Autoregressive flows [71]. However, none of these approaches were investigated for disentanglement. Closer to our work, Ma *et al.* [107] decouple the global and local representation of images by embedding a NF in the VAE framework to model

the decoder. This shows that cNF naturally tends to disentangle, hence justifying the approach taken in this work.

Following the work of Locatello *et al.* [103], a handful of methods have recently tried to propose weak-supervision methods based on pairs of images [104] to ensure the representation is identifiable. This was extended to causal variables in [16] and to sequences in [100]. However, those methods rely either on knowing the number of changing variables or on labels.

Contributions: Our framework seeks to solve the limitations of previous works by first proposing to *formalize the definition of the factors* and then building on this formalism to *introduce a new model that accounts for the potential dependency between the variables*. Our model extends the classical VAE approach through the use of additional Normalizing Flows. In stark contrast to previous works, we show that, by leveraging the data structure through a simple shuffle operation, we can obtain a *provably disentangled representation without any additional loss or supervision* (no data augmentation [7], contrastive loss [112], external models [180] or adversarial training [56]), making our method undemanding and modality free, while having theoretical guarantees. This is detailed in the following sections.

3.3 Problem Formulation

Before trying to disentangle the static and dynamic factors, we propose a new definition of those concepts. This is done through formalizing the video generative process as a latent variable model that comprises two disjoint codes representing the time varying and time invariant factors. We further provide the assumptions underlying this partition.

Disentangled video generative model. We propose a model where the generative process of each frame consists in first drawing a latent code $\boldsymbol{z}_t$ from its probability density $p(\boldsymbol{z}_t)$ and then obtaining the observations $\boldsymbol{x}_t$ by sampling from $p(\boldsymbol{x}_t|\boldsymbol{z}_t)$. More formally, the frame generative process is:

$$\boldsymbol{z}_t \sim p(\boldsymbol{z}_t), \qquad \boldsymbol{x}_t = \boldsymbol{g}(\boldsymbol{z}_t) \tag{3.3}$$

where $p(\boldsymbol{z}_t)$ is a smooth, continuous density on $\mathcal{Z}$ with full support and $\boldsymbol{g} : \mathcal{Z} \to \mathcal{X}$ is a smooth and invertible function with smooth inverse (diffeomorphism) mapping the frame representation space $\mathcal{Z} \subseteq \mathbb{R}^k$ to the observation space $\mathcal{X} \subseteq \mathbb{R}^n$, with usually $n \gg k$.

Based on the static/dynamic disentanglement assumption presented in Sec.3.2, the video sequence frame codes $\boldsymbol{z}_{1:T}$ can be partitioned into a single invariant static code $\boldsymbol{s}$ and temporally varying dynamic codes $\boldsymbol{d}_{1:T}$, i.e., $\mathcal{Z} = \mathcal{S} \times \mathcal{D}$ with $\mathcal{S} \subseteq \mathbb{R}^{n_s}$ and $\mathcal{D} \subseteq \mathbb{R}^{n_d}$, $k = n_s + n_d$:

$$\boldsymbol{z}_{1:T} = (\boldsymbol{s}, \boldsymbol{d}_{1:T}) \sim p(\boldsymbol{s})p(\boldsymbol{d}_1|\boldsymbol{s}) \prod_{t=2}^{T} p(\boldsymbol{d}_t|\boldsymbol{s}, \boldsymbol{d}_{<t}), \tag{3.4}$$

$$\boldsymbol{x}_t = \boldsymbol{g}(\boldsymbol{s}, \boldsymbol{d}_t), \tag{3.5}$$

with $\boldsymbol{s}$ corresponding to the first n_s dimensions of $\boldsymbol{z}_t$, $\boldsymbol{s} = \boldsymbol{z}_{t,1:n_s}$. Each frame is generated from its corresponding static and dynamic factors with the dynamic factors obtained given previous time steps **and** the static factors $\boldsymbol{s}$. From Eq.3.4, it is directly given that $\forall \boldsymbol{x}_{1:T}$ and $t, t' \in \{1, \cdots, T\}$, $\boldsymbol{g}^{-1}(\boldsymbol{x}_t)_{1:n_s} = \boldsymbol{g}^{-1}(\boldsymbol{x}_{t'})_{1:n_s} = \boldsymbol{s}$. That is the static factors are invariant for each sequence. As a result, it can be noted that this expression actually matches the formalism presented in the context of style/content disentanglement [151] for static images but generalized here to multiple temporally varying views. It ensues that the dynamic factors can be defined based on the same assumptions than the ones adopted for the 'style' factors in [151]:

1. $p(\boldsymbol{d}_t|\boldsymbol{s}, \boldsymbol{d}_{<t}) = \delta(\boldsymbol{d}_{t,A_t^c} - \boldsymbol{d}_{t-1,A_t^c}) p(\boldsymbol{d}_{t,A_t}|\boldsymbol{s}, \boldsymbol{d}_{<t})$ with $A_t \subseteq \{1, ...n_d\}$ denoting the subset of varying dynamic components at time step $t \geq 2$ with associated $p(A_t)$, and A_t^c the complement of A_t.

2. $\forall i \in \{1, ...n_d\}$, $\exists t \in \{2, ...T\}$, $A_t \subseteq \{1, ...n_d\}$ s.t. $i \in A_t$; $p_{A_t}(A_t) > 0$ and for any $(\boldsymbol{s}, \boldsymbol{d}_{<t})$, $p(\boldsymbol{d}_{t,A_t}|\boldsymbol{s}, \boldsymbol{d}_{<t})$ is smooth and fully supported in some open non empty subset containing $\boldsymbol{d}_{t-1,A_t}$.

The second assumption means that every dynamic variable varies with time. However, as stated by the first assumption, it does not force all dynamic components to change at each time step or for each sequence, echoing the fact that only a subset of the dynamic factors might be fluctuating in a given sequence.

This video generative model does not make any assumption on the relationship between the factors and allows for arbitrary marginals $p(\boldsymbol{s}, \boldsymbol{d}_{1:T})$ with potentially complex non trivial statistical relationships between (and within) $\boldsymbol{s}$ and $\boldsymbol{d}_{1:T}$. That is $I(\boldsymbol{s}, \boldsymbol{d}_{1:T}) \geq 0$. Overall, this illustrates that, while the static and dynamic factors constitute two separate semantic concepts, they might not be independent. As a result, similarly to [151, 104] for pairs of images, we propose to further extend our description using a Structural Causal Model (SCM) [168] to express the relation between factors.

Causal model. From the assumption that $\boldsymbol{s}$ is invariant throughout the sequences while the $\boldsymbol{d}_{1:T}$ varies, it results that the static factors may causally influence the dynamic ones but not the opposite. Indeed, if $\boldsymbol{s}$ would depend on some dynamic variable, following the definition of $\boldsymbol{d}$ above, this dynamic variable varies and therefore $\boldsymbol{s}$ would vary as well, contradicting the static invariance. Assuming the dynamic variables undergo temporally coherent perturbations, videos can thus be interpreted as temporally intervened sequences where successive frames result from the causal generative model under soft intervention on the dynamic factors $\boldsymbol{d}$. This can be formally written as :

$$\boldsymbol{s} = \boldsymbol{h}_s(\boldsymbol{\epsilon}_s), \qquad \boldsymbol{d}_1 = \boldsymbol{h}_d(\boldsymbol{s}, \boldsymbol{\epsilon}_{d,1}), \qquad \boldsymbol{d}_t = \boldsymbol{h}'_d(\boldsymbol{s}, \boldsymbol{\epsilon}_{d,<t}, \boldsymbol{\epsilon}_{d,t}) \tag{3.6}$$

with $\boldsymbol{\epsilon}_s$ and $\boldsymbol{\epsilon}_{d,1:T}$ independent exogenous variables and $\boldsymbol{h}_s$,$\boldsymbol{h}_d$,$\boldsymbol{h}'_d$ functions describing the causal mechanism. Eq.3.6 shows that $\boldsymbol{s}$ might causally influence the specific values of the dynamic factors but also their possible variations.

By formalizing the video generative process based on static/dynamic factors disentanglement, the link with data augmentation generative process [151] becomes apparent. In the particular case where T=2, the two problems are equivalent with the exception that, compared to style/content, the augmentation is not human-designed but is a result of temporal variations in the sequence. This analogy provides an interpretation of videos as augmented views of a scene through temporal augmentation between each frame. Moreover, in [151] an approach is proposed to provably identify the content for style/content disentanglement. The similarities between the two problems thus also hint at the possibility of proposing a similar approach that would provably isolate the static **and** dynamic factors for sequential disentanglement. We formulate now this idea of provably identifying the different factors.

Definition 3.1. The ground truth factors $\boldsymbol{s}$ and $\boldsymbol{\epsilon}_{d,1:T}$ are said to be identified by learned vectors $\boldsymbol{f}$ and $\boldsymbol{\lambda}_{1:T}$ if there exist invertible transformations $\boldsymbol{m}$ and $\boldsymbol{a}$ s.t. $\boldsymbol{f} = \boldsymbol{m}(\boldsymbol{s})$, $\boldsymbol{\lambda}_{1:T} = \boldsymbol{a}(\boldsymbol{\epsilon}_{d,1:T})$ i.e., $\boldsymbol{f}$ and $\boldsymbol{\lambda}_{1:T}$ are reparametrizations of $\boldsymbol{s}$, $\boldsymbol{\epsilon}_{d,1:T}$ respectively.

This means that the learned codes $\boldsymbol{f}$ should capture exactly the static factors $\boldsymbol{s}$, while $\boldsymbol{\lambda}_{1:T}$ should encode the dynamic conditionally to $\boldsymbol{s}$. This is termed as 'conditional disentanglement'. The use of separate notations

serves to differentiate the learned codes $\boldsymbol{f}$ and $\boldsymbol{\lambda}_{1:T}$ (extracted by the designed network) from the ground truth factors $\boldsymbol{s}$ and $\boldsymbol{\epsilon}_{d,1:T}$ (imposed by the video generative process) which might not be equal in practice. The potential mismatch between the objective and the generative process is avoided by representing the exogeneous dynamic variables $\boldsymbol{\epsilon}_{d,1:T}$ instead of $\boldsymbol{d}_{1:T}$. If the static/dynamic factors are independent, $\boldsymbol{\epsilon}_{d,1:T}$ and $\boldsymbol{d}_{1:T}$ are equivalent and our formulation becomes therefore a generalization of the objective of Sec.3.2 that also accounts for causal dependencies between factors.

3.4 Method

We seek to learn from the video sequences $\boldsymbol{x}_{1:T}$, two latent codes, a static code $\boldsymbol{f} \in \mathbb{R}^{n_f}$ and dynamic codes $\boldsymbol{\lambda}_{1:T} \in \mathbb{R}^{n_\lambda}$ such that these codes identify the ground-truth factors as presented in Def.1. We start by deriving from our formalism sufficient conditions to provably extract the factors. We then propose a practical model and show how, strictly by leveraging the data structure, this method does achieve the disentanglement conditions. The complete model is summarized in Fig.3.3.

3.4.1 Sufficient Conditions for Disentanglement

Static disentanglement. The objective is to learn an encoder posterior distribution $q(\boldsymbol{f}|\boldsymbol{x}_{1:T})$ mapping the sequences to their static code such that the static codes $\boldsymbol{f}$ sampled from $q(\boldsymbol{f}|\boldsymbol{x}_{1:T})$ solely capture the static factors $\boldsymbol{s}$.

Proposition 3.1. *Consider the generative process and conditions presented in Sec.3.3. Further assume that $dim(\boldsymbol{f}) = n_f \geq n_s$. Let $\boldsymbol{x}$ and $\boldsymbol{x}^*$ denote two frames (or disjoint sets of frames) belonging to the same sequence $\boldsymbol{x}_{1:T}$ and $\mathbb{D}$ be a divergence between two distributions. Given unlimited data from $p(\boldsymbol{x}_{1:T})$, any candidate posterior distribution $q(\boldsymbol{f}|\boldsymbol{x}_{1:T})$ that minimizes*

$$L_{prop1} = \mathbb{E}_{p(\boldsymbol{x}_{1:T})}\mathbb{D}(q(\boldsymbol{f}|\boldsymbol{x}), q(\boldsymbol{f}|\boldsymbol{x}^*)) - I(\boldsymbol{f}, \boldsymbol{s}) + I(\boldsymbol{s}, \boldsymbol{s}), \tag{3.7}$$

is disentangled in the sense that $q(\boldsymbol{f}) = \int q(\boldsymbol{f}|\boldsymbol{x}_{1:T})p(\boldsymbol{x}_{1:T})d\boldsymbol{x}_{1:T}$, the aggregated posterior, is a reparametrization of $p(\boldsymbol{s})$.

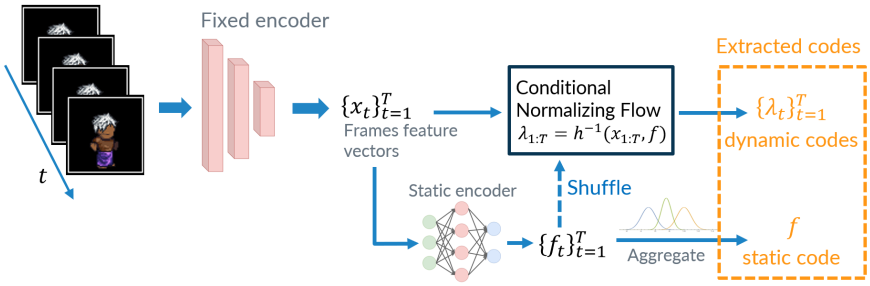


Fig. 3.3 Schematic view of the proposed model. A pretrained Convolutional encoder embeds each frame separately into the frame latent space $\boldsymbol{x}_{1:T}$. These vectors are used as input to the static encoder which estimates the static codes from each individual frame. The estimations are then aggregated giving the static code $\boldsymbol{f}$. The static vectors serve to condition a Conditional Normalizing Flow that models the likelihood of the frame feature vectors $\boldsymbol{x}_{1:T}$. The transformed vectors $\lambda_{1:T}$ correspond to the dynamic codes. The model is trained using the loss $\mathcal{L}$ in Eq.3.12.

The proof is provided in App.A.1.2. Here $\mathbb{D}$ can be any divergence between distributions, i.e. a non-negative functional that vanishes iff its two arguments coincide, such as the Kullback–Leibler divergence. Intuitively, by imposing $\boldsymbol{f}$ to be both invariant to the dynamic factors and informative of the static ones, the code $\boldsymbol{f}$ will capture all and only $\boldsymbol{s}$. Prop.3.1 extends Thm.4.4 of [151] to distributions and cope with an unknown number of static variables n_s. Note that the invariance term is essential: simply maximizing $I(\boldsymbol{x}_{1:T}, \boldsymbol{f})$ as proposed in Sec.3.2 is not sufficient to extract $\boldsymbol{s}$. Indeed, as shown in [48, 171], a global aggregated code $\boldsymbol{f}$ could encode the dynamic information and should therefore additionally be made invariant.

Dynamic disentanglement. As explained in Sec.3.3, because the static and dynamic factors might not be independent, the goal for the dynamic codes $\boldsymbol{\lambda}_{1:T}$ is to capture $\boldsymbol{\epsilon}_{d,1:T}$. Unfortunately, following an approach similar to Prop.3.1 is not possible in this case. Indeed, this approach would require to have access to two samples $\boldsymbol{x}$ and $\boldsymbol{x}^*$ that share the same motion but have different static factors. However, by the definition of the dynamic factors (Sec.3.3) there are no easily obtainable such samples since $\boldsymbol{d}_t$ is specific to each frame and varies at unknown time steps. A solution proposed in [7] consists in generating positive samples using static data augmentation. However, designing those augmentations might be non trivial and dataset specific, with the risk of artificially and indirectly defining what is static/dynamic through the chosen augmentations. Moreover, only the

static part should be affected, limiting the amount of available augmentations, possibly favoring static information in the dynamic codes. Instead, an alternative approach is proposed here that does not require positive samples.

Proposition 3.2. *Consider the same framework as in Prop.3.1 and the static code $\boldsymbol{f}$ to be disentangled as described above. Further assume that $dim(\boldsymbol{\lambda}) = n_\lambda \geq n_d$. Any smooth and invertible candidate function $\boldsymbol{h}$ s.t. $\boldsymbol{x}_{1:T} = \boldsymbol{h}(\boldsymbol{\lambda}_{1:T}, \boldsymbol{f})$ which minimizes*

$$L_{prop2} = I(\boldsymbol{f}, \boldsymbol{\lambda}_{1:T}), \tag{3.8}$$

is disentangled in the sense that $q(\boldsymbol{\lambda}_{1:T})$ is a reparametrization of $p(\boldsymbol{\epsilon}_{d,1:T})$.

The proof is provided in App.A.1.3. Intuitively, assuming that $\boldsymbol{f}$ represents exactly $\boldsymbol{s}$ thanks to Prop.3.1 and that both $\boldsymbol{f}$ and $\boldsymbol{\lambda}_{1:T}$ are not redundant, because $\boldsymbol{h}$ is invertible, the dynamic codes $\boldsymbol{\lambda}_{1:T}$ will have to be informative and to capture all and only $\boldsymbol{\epsilon}_{d,1:T}$.

Overall, by using the formalism in Sec.3.3, we are able to discover sufficient conditions such that the learned codes are reparametrizations of the ground truth factors. We will now show how we learn $q(\boldsymbol{f}|\boldsymbol{x}_{1:T})$ and $\boldsymbol{h}$ in practice so that the conditions are enforced from the observed data only, without requiring explicit formulations of L_{prop1} and L_{prop2}.

3.4.2 Practical Implementation

Compared to previous methods that solely rely on VAEs, we propose instead to improve sequence modeling based on additional Normalizing Flows (NF) [130]. A NF is a transformation of a random variable through a sequence of differentiable invertible mappings. As a result, it is a natural candidate to learn the invertible function $\boldsymbol{h}$ required in Prop.3.2. More specifically, assuming the common static code $\boldsymbol{f}$ has been extracted from the frames, we model the frames likelihood $p(\boldsymbol{x}_{1:T}|\boldsymbol{f})$ using a cNF [161] as

$$p(\boldsymbol{x}_{1:T}|\boldsymbol{f}) = p(\boldsymbol{\lambda}_{1:T}|\boldsymbol{f}) \left| \det \frac{\delta \boldsymbol{h}^{-1}(\boldsymbol{x}_{1:T}, \boldsymbol{f})}{\delta \boldsymbol{x}_{1:T}} \right| \tag{3.9}$$

with $\boldsymbol{\lambda}_{1:T} = \boldsymbol{h}^{-1}(\boldsymbol{x}_{1:T}, \boldsymbol{f})$ the dynamic codes. This term can be evaluated through exact likelihood evaluation. Because a NF is a bijective transformation and $\boldsymbol{f}$ is common to all frames, the transformed latents $\boldsymbol{\lambda}_{1:T}$ will retain and encode the dynamics of the sequence. Since the NF is condi-

tioned on the static code $\boldsymbol{f}$, the dynamics are encoded conditionally to the static part. The NF is based on affine coupling [35] and LSTM layers that model the temporal behavior of the sequence, similarly to [108], and provide a bias towards encoding the dynamics. Because, for a disentangled model, the transformation $\boldsymbol{h}$ needs to fully model the conditional dependence and causal mechanism, the final distribution $p(\boldsymbol{\lambda}_{1:T}|\boldsymbol{f})$ is replaced with $p(\boldsymbol{\lambda}_{1:T})$. Details on the architecture are provided in App.A.2.3.

It remains to show how to extract the static code $\boldsymbol{f}$ a priori before solving Eq.(3.9). Variational inference is used in order to learn an approximate posterior $q(\boldsymbol{f}|\boldsymbol{x}_{1:T})$. Following the standard VAE framework, the marginal data log-likelihood is lower estimated by the Evidence Lower BOund (ELBO):

$$\log p(\boldsymbol{x}_{1:T}) \geq \mathbb{E}_{q(\boldsymbol{f}|\boldsymbol{x}_{1:T})} \log p(\boldsymbol{x}_{1:T}|\boldsymbol{f}) - KL(q(\boldsymbol{f}|\boldsymbol{x}_{1:T})||p(\boldsymbol{f})) \tag{3.10}$$

$$\geq \mathbb{E}_{q(\boldsymbol{f}|\boldsymbol{x}_{1:T})} \log p(\boldsymbol{\lambda}_{1:T}) + \log \left| \det \frac{\delta \boldsymbol{h}^{-1}}{\delta \boldsymbol{x}_{1:T}} \right| - KL(q(\boldsymbol{f}|\boldsymbol{x}_{1:T})||p(\boldsymbol{f})) \tag{3.11}$$

with the two first terms corresponding to the cNF likelihood. The detailed derivation is shown in App.A.1.1. Overall, the model can be described as a VAE that extracts a single latent code $\boldsymbol{f}$ from the observations and has a cNF decoder to model the likelihood of the frames. By maximizing the ELBO, we learn both an approximation of the encoder posterior $q(\boldsymbol{f}|\boldsymbol{x}_{1:T})$ and the cNF likelihood i.e., the transformation $\boldsymbol{h}$. This model does not make any assumption on the independence between the codes and encompasses the potential conditional dependencies through the cNF. In comparison, the 'full q' model (Sec.3.2) is equivalent to model $p(\boldsymbol{x}_{1:T}|\boldsymbol{f})$ using a VAE with a latent vector for each time step and a simple concatenation of the codes as inputs to the generative network. The use of a NF acts as a more flexible and expressive 'mix-up' of the two extracted codes.

The prior distribution $p(\boldsymbol{f})$ is further modeled using a NF. This is in contrast with most works that propose to use NF to generate more flexible posteriors instead of priors [90, 147]. As demonstrated in [29], if these two approaches are equivalent during training, at test time, for sampling, a NF prior allows to have a longer decoder path and produces better samples. To model the inference posterior distribution, the static code encoder $q(\boldsymbol{f}|\boldsymbol{x}_{1:T})$ needs to output a single latent code from multiple frames. This is achieved by means of an aggregated posterior distribution [14]. The

static code distribution $q(\boldsymbol{f}|\boldsymbol{x}_{1:T})$ is constructed as the product of individual distributions estimated from each frame by a static encoder network: $q(\boldsymbol{f}|\boldsymbol{x}_{1:T}) = \prod_{t=1}^{T} q(\boldsymbol{f}_t|\boldsymbol{x}_t)$. Compared to previous works where $\boldsymbol{f}$ is inferred through recurrent layers, the aggregated posterior is inherently permutation invariant, which provides a bias towards our disentanglement objective.

Shuffle constraint. Unfortunately, the numerous disentanglement biases included in this proposed architecture might not be sufficient to achieve a well-disentangled model. In order to achieve a clean disentanglement, it is therefore proposed to leverage the sequential nature of video data directly into the framework. Indeed, thanks to the aggregated posterior, the model directly provides the static code estimation for individual frames $q(\boldsymbol{f}_t|\boldsymbol{x}_t)$, without any additional computation. The idea is to condition the cNF, during training, on the shuffled $\boldsymbol{f}_{1:T}$ before aggregation, instead of conditioning the cNF on the aggregated static code $\boldsymbol{f}$. This is justified by the fact that, for a disentangled model, the $\boldsymbol{f}_{1:T}$ need to be invariant i.e., $p(\boldsymbol{x}_{1:T}|\boldsymbol{f}) = p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{\pi(T)})$ with $\pi(T)$ denoting a random permutation. Using this shuffled operation, the loss becomes:

$$\boxed{\mathcal{L} = \mathbb{E}_{p(\boldsymbol{x}_{1:T})}\mathbb{E}_{q(\boldsymbol{f}|\boldsymbol{x}_{1:T})} - \alpha \log p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{\pi(T)}) + \beta KL(q(\boldsymbol{f}|\boldsymbol{x}_{1:T})||p(\boldsymbol{f}))} \tag{3.12}$$

which is the negative dataset ELBO where $p(\boldsymbol{x}_{1:T}|\boldsymbol{f})$ has been replaced by $p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{\pi(T)})$. Following the method of β-VAE [60], the different terms in $\mathcal{L}$ have additionaly been weighted using weight coefficients α and β.

Proposition 3.3. *Consider any $\boldsymbol{h}$ and $q(\boldsymbol{f}|\boldsymbol{x}_{1:T})$ as described above which minimize $\mathcal{L}$ and further assume $\alpha \gg \beta$. Minimizing $\mathcal{L}$ is equivalent to minimizing both L_{prop1} and L_{propr2}. Prop.3.1 and 3.2 are satisfied and the learned representation $(\boldsymbol{f}, \boldsymbol{\lambda}_{1:T})$ is disentangled in the sense of Def.3.1.*

The proof is given in App.A.1.4. Intuitively, shuffling the static codes $\boldsymbol{f}_{1:T}$ and using them to reconstruct the frames, enforces the static code to be time invariant, while simultaneously being informative due to the reconstruction objective (favored by imposing $\alpha \gg \beta$). Meanwhile, because we minimize $-\log p(\lambda_{1:T}) \approx H(\lambda_{1:T})$ through the cNF, the dynamic code is encouraged to be non informative. In turn, this leads to the minimization of both L_{prop1} and L_{prop2}, resulting in a disentangled model. Further details are provided in App.A.3.2.

Summary: we showed that, through a simple shuffle of the static codes, the proposed model is provably disentangled without requiring any additional complex loss, unlike previous methods, with all disentanglement conditions directly satisfied by our novel evidence lower bound. This makes our method modality free and highly efficient by avoiding any unnecessary constraint. In the case where the factors are independent, the model tries to capture both $\boldsymbol{s}$ and $\boldsymbol{d}_{1:T}$. However, the proposed model also generalizes to factorize the static and dynamic factors without assuming independence. It does so by directly modeling the causality and only capturing $\boldsymbol{\epsilon}_{d,1:T}$.

Using pretrained Autoencoders. Similarly to [100], to facilitate disentanglement, instead of directly extracting the two separate codes, it is proposed to first learn an Autoencoder trained to encode/decode the frames without disentanglement. Gaussian noise is added on the frame feature vectors to prevent collapse. After convergence, the autoencoder parameters are frozen and we learn the static encoder and conditional Normalizing Flow which provide the disentangled representation from the entangled one. Because a Normalizing Flow is invertible, it ensures no information is lost and the frozen decoder can be used to recover the frames without fine-tuning. This permits to fully factorize the disentanglement problem from the task of mapping high-dimensional complex images into a low dimensional space.

3.5 Experiments

In order to validate the effectiveness of the proposed model, it is compared qualitatively and quantitatively to state-of-the-art sequence disentanglement learning methods including MoCoGAN [146], DSVAE [99], R-WAE [56], S3VAE [180], SKD [10], CDSVAE [7] and SPYL [112].

Datasets. Similarly to the baseline works, the model is evaluated using the standard *LPCSprites* [128] and *MUG* [3] databases, which contain videos of animated cartoon characters and facial expressions respectively. As all methods manage to provide high quality results on the original *LPCSprites* dataset [112], it is modified in order to include more complex characters with static factors other than colors. Additionally, the *MPI3D* database [51] is also considered in the experiments. It consists of a real-

world robot arm carrying various objects in different positions with the arm motion made dependent on the background light, and the speed of the arm dependent on the object size. We further evaluate on the human action recognition *MHAD* dataset [25] comprising videos of humans performing diverse actions. Finally, based on the *dSprites* dataset [109], sequences of moving 2D shapes are generated. Two variants are proposed, one with fixed rotation and all factors independent (*s-dSprites*) and the other with rotating shapes and objects that bounce on the edges according to their size (*c-dSprites*).

Disentanglement metrics. Quantitative evaluation is provided following the common benchmark for sequential disentanglement [99, 7, 180]. After encoding a sequence with known factors, either the static or dynamic code is replaced by a sample from its prior before reconstruction. A classifier, pretrained with full supervision, is then used to measure how well the factors associated with the fixed code are preserved for the reconstructed sequence. Four metrics are extracted : the accuracy for the fixed factors (Acc), the intra-entropy ($H(y|x)$) that measures the classifier confidence for the predictions, the inter-entropy ($H(y)$) that estimates the diversity for the sampled factors and the inception score (IS) that measures the generator performance.

Details on the model architecture and hyperparameters, datasets, disentanglement metrics as well as additional results can be found in App.A.2 and A.3.

Results. The *MUG*, *MHAD* and *s-dSprites* datasets have independent static-dynamic factors. Their corresponding results are shown in Tab.3.1. As expected, when the factors are independent all recent methods manage to provide a disentangled representation. In particular, while using a simple shuffle operation, our proposed model performs on par with the complex state-of-the-art methods on all datasets. Despite the dynamic codes being extracted conditionally to f, the learned representation correctly factorizes the independent factors. Only the SPYL method [112] gives slightly better results on *MUG* and *MHAD*. However, we show in App.A.3.1 that this can be mainly explained by the tendency of SPYL to encode too much information in its dynamic codes. In Fig.3.2, qualitative results for the MUG dataset are displayed with sequences generated by swapping the dynamic and static codes of test samples. The resulting videos accurately retain the person identities while exchanging the face motions with high quality non blurry outputs.

Table 3.1 Disentanglement metrics on *s-dSprites*, *MUG* and *MHAD* datasets with independent factors. The dynamic factors are highligted in red.

MUG	Acc↑ (%) action	IS↑	$H(y\|x)$↓	$H(y)$↑
MoCoGAN	63.12	4.332	0.183	1.721
DSVAE	54.29	3.608	0.374	1.657
R-WAE	71.25	5.149	0.131	1.771
S3VAE	70.51	5.136	0.135	1.760
SKD	77.45	**5.569**	**0.052**	1.769
C-DSVAE	81.16	5.341	0.092	1.775
SPYL	**85.71**	5.548	0.066	**1.779**
Ours	84.32	5.329	0.068	1.778

s-dSprites	Acc↑ (%) color	shape	size	IS↑
DSVAE	79.38	42.10	43.61	3.035
CDSVAE	96.18	98.22	98.71	3.995
SPYL	97.76	98.72	98.12	3.685
Ours	**98.91**	**98.98**	**98.76**	**4.263**

MHAD	Acc↑ (%) action	IS↑	$H(y\|x)$↓	$H(y)$↑
C-DSVAE	65.59	1.749	0.420	1.282
SPYL	**68.63**	1.780	**0.362**	1.299
Ours	67.65	**1.896**	0.371	**1.325**

For the *LPCSprites*, *c-dSprites* and *MPI3D* datasets the static-dynamic factors are not independent. As an example, the *c-dSprites* dataset displays 2D objects that move and rotate. Because the 2D shapes are either squares, ellipses or hearts, even if the rotation is generated independently, it is still causally influenced by the shape due to the different order of rotational symmetry between the shapes. Overall, this example highlights the fact that, even in simple cases, the static content of a scene may influence the dynamic of the sequence by conditioning the possible variations between frames. In Tab.3.2 the results for the three aforementioned datasets are given. Compared to the case with independent factors, the performance of the baseline methods significantly drop and those models fail to provide a disentangled representation. As explained in Sec.3.2, these methods ex-

Table 3.2 Disentanglement metrics on *c-dSprites*, *MPI3D* and *LPCSprites* with dependent factors. In red and blue are highligted respectively the dynamic factors and the static factors that condition the motion.

c-dSprites	Acc↑ (%) color	shape	size	IS↑	$H(y\|x)$↓	$H(y)$↑
DSVAE	72.09	32.66	79.38	4.128	0.046	1.41
CDSVAE	96.33	74.32	93.17	4.141	0.028	1.39
SPYL	68.41	33.28	75.10	**4.392**	0.027	1.43
Ours	**98.92**	**98.98**	**98.59**	4.371	**0.008**	**1.44**

MPI3D	Acc↑ (%) shape	size	camera	light	IS↑	$H(y\|x)$↓	$H(y)$↑
DSVAE	28.29	65.57	98.81	87.71	3.505	0.056	1.19
CDSVAE	43.81	79.17	**99.99**	**99.94**	3.487	0.138	1.26
SPYL	16.78	50.29	33.32	86.54	3.819	0.043	1.28
Ours	**90.15**	**97.56**	99.97	99.93	**3.849**	**0.029**	**1.29**

LPCSprites	Acc↑ (%) body	hair	orient.	action	IS↑	$H(y\|x)$↓	$H(y)$↑
CDSVAE	54.94	85.75	33.32	67.71	2.737	1.6e-3	1.00
SPYL	70.89	99.89	33.32	**99.86**	2.735	2.0e-3	**1.01**
Ours	**99.97**	**99.91**	**99.94**	99.83	**2.768**	**3.3e-4**	**1.01**

tract the static and dynamic variables separately and enforce independence between the codes. This conflicts with the inherent causality of the ground truth data generative process (Sec.3.3) and results in ambiguity into where the shared information between the dynamic and static factors should lie. If the dynamic code is extracted independently, it will inevitably have to retain some static information, translating in entangled codes and potentially encouraging static factors into the dynamic variables. In contrast, our method manages to provide a disentangled representation and generalizes to settings with complex statistical relationships between the factors. The conditionality is properly captured and high accuracy scores are maintained. These results are illustrated in Fig.3.1 which shows sequences generated by fixing the static code and sampling the dynamic one. Compared to the competing CDSVAE model [7], the designed method presents improved sample quality that maintains the object shapes and identity. The samples display diverse motions that abide by the causal mechanism (i.e.,

only vertical motion for the *MPI3D* sample and fixed orientation for the *LPCSprites* one, see App.A.2.1 for details on the datasets).

Limitations. The proposed formalism (Sec.3.3) defines the factors. But, by doing so, it also draws the intrinsic limitations of the static/dynamic disentanglement problem as it provides what could be disentangled at best through DSVAE based methods. First, it is the dataset itself that defines what is static/dynamic, which might not match the human intuition or expectations, e.g., for the *LPCSprites* dataset in Fig.3.1, from a human perspective the orientation of a character would probably be labelled as a motion of that specific character identity. However, from the data perspective the orientation is static since it is time invariant. Second, the static code can only capture what is always invariant, i.e., factors that are static in every sequence of the dataset. A factor that can be either static or dynamic, depending on the sequence, is encompassed in $\boldsymbol{d}_{1:T}$. Hence, future work may investigate whether the proposed method can be extended to adaptively extract the static code for each sequence and further disentangle the factors at the sequence level.

3.6 Conclusion

In this work, we introduce a novel formal approach toward sequential data disentanglement. Diverging from previous works, it extends the VAE framework with a conditional Normalizing Flow that directly models the relationships between the static and dynamic factors, making it closer to the data generative process. This constitutes a crucial change compared to the baseline methods by lifting their most detrimental assumption. It translates into large improvements over the state-of-the-art models, which fail to disentangle dependent factors even in simple cases. To further enforce disentanglement, a simple shuffle constraint is proposed which leads to a novel ELBO formulation and our model to be provably disentangled. Hence, compared to the baselines, our method is simple, non dataset or domain specific and theoretically justified. It proved its ability to properly disentangle the static/dynamic factors on multiple datasets and to generalize to less restrictive cases with dependent variables. Interestingly, while our method has been only evaluated for videos, the presented approach directly extend to other domains such as audio or biology which could be explored in future works.

3.7 Beyond the Framework: Two Emerging Research Directions

The framework developed in this chapter achieves its stated objective: a provably disentangled representation of sequential data that separates static and dynamic factors, even when those factors are statistically dependent. Yet in reaching this goal we have, perhaps inevitably, opened new questions that extend well beyond the sequential setting. Two directions stand out as particularly natural and compelling, and both are taken up in the chapters that follow.

Question 1. *Can the conceptual machinery developed here, founded on* functional *rather than statistical independence, be generalised to disentangle more than two factors simultaneously, and can such a generalisation be carried out in a fully unsupervised manner on static data?*

Question 2. *What happens if the disentanglement constraint is removed entirely: can the same generative structure, without any disentanglement requirement, be turned into a powerful and general-purpose video representation?*

We briefly outline the motivating ideas behind each question in turn.

3.7.1 Direction 1 — Multi-Factor Disentanglement on Static Data

Throughout this chapter the disentanglement problem has been formulated as a binary decomposition: a single static factor $\boldsymbol{s}$ and a dynamic counterpart $\boldsymbol{d}_{1:T}$. A natural question is whether the core ideas can be extended to separate an arbitrary number of factors simultaneously. Consider, for instance, a dataset of facial images. A human observer would readily identify several distinct, independently controllable attributes: hair length, eye colour, head orientation, skin tone, and so on, that together constitute the meaningful degrees of freedom of the data. Can one recover those factors automatically, without supervision, and without assuming they are statistically independent?

This question connects to the classical problem of nonlinear Independent Component Analysis (ICA), where the standard strategy is to look for a representation whose components are statistically independent [96, 113, 88, 60, 103]. As the analysis of the present chapter makes clear, however, statistical independence is rarely a faithful proxy for the factors that align with human intuition. Hair length and gender are semantically distinct

concepts yet they are statistically correlated; altitude and temperature are separately controllable physical quantities yet they tend to co-vary in natural data. Forcing independence on the latent space therefore risks misidentifying the true generative factors or conflating them in ways that break semantic interpretability.

A key insight of the present work is that what *really* distinguishes the static factor $\boldsymbol{s}$ from the dynamic factors $\boldsymbol{d}_{1:T}$ is not their statistical relationship but the way they act on the data: each factor influences the observations through a distinct, separable mechanism that can be varied while leaving the other factor's contribution unchanged. This is precisely the intuition captured by the shuffle constrain: by intervening on the temporal ordering we verify that $\boldsymbol{s}$ can be held fixed while $\boldsymbol{d}$ varies, and vice versa. In other words, what defines a factor is not statistical orthogonality but *functional* orthogonality: the ability to act on one degree of freedom independently of the others.

Extending this idea to the multi-factor static setting is conceptually straightforward but technically non-trivial.One could imagine constructing a dataset in which pairs of observations differ in exactly one factor, analogous to the pairs induced by the shuffle operation in the sequential case, and applying the same conditional normalizing flow framework generalized to multiple conditioning variables. Such an approach has been explored in the literature [104, 16, 100], and the parallel with our framework is instructive. The fundamental difficulty, however, is one of data: interventional datasets that provide pairs of observations where a precise subset of factors has been changed are *extremely* rare in practice. Curating them requires controlled experimental conditions, fine-grained annotation, or synthetic data generation, none of which are available at scale for most real-world domains. This is, in fact, precisely the problem that the shuffle trick circumvents in the sequential case: temporal structure provides a natural and freely available source of such interventions.

For static data, no such structure exists by definition. Yet a human observer does not require paired observations to identify the factors of variation in a dataset of faces: the factors are somehow visible *in the data itself*, through the pattern of variation across the collection of images. This observation suggests that interventional pairs may not be strictly necessary, and that the information required for disentanglement is already implicitly present in an unpaired dataset, provided the dataset is rich enough to exhibit all combinations of the relevant factors.

This is the starting point for Chapter 4. Rather than relying on paired

data or causal assumptions, we redefine what it means for a latent factor to be a meaningful concept. Concretely, we formalize the notion that factors act *independently on the data* through an orthogonality condition on the columns of the Jacobian of the generative mapping: a local, functional criterion that captures the same intuition as the shuffle constraint but applies to static data without requiring pairs. This condition is closely related to the Independent Causal Mechanisms (ICM) principle [134], but it is deliberately formulated without any causal commitment: we do not assume a particular causal graph, only that the generative mechanisms are locally decoupled. Crucially, we show that this local orthogonality condition, combined with a global completeness assumption stating that the dataset covers all combinations of the latent factors (Asm.4.2), is *sufficient* for identifiability. In this sense, the unpaired dataset itself plays the role of the interventional pairs: if every combination of factors is represented in the data, then the interventional information is implicitly available, and disentanglement becomes provably achievable in a fully unsupervised manner.

The result is a new framework for multi-factor disentanglement that is neither independence-based nor causality-based, but rather grounded in the functional geometry of the generative process. It extends the spirit of the present chapter, replace statistical assumptions with functional ones, to a setting where the data is static, the number of factors is arbitrary, and no supervision or paired observations are required.

3.7.2 Direction 2 — Unconstrained Global Representations for Video

The second direction emerges from a closer examination of the role played by the shuffle constraint in our framework. Recall that the global latent variable $\boldsymbol{f}$ is *encouraged* to encode the static content of a sequence $\boldsymbol{s}$ precisely because the shuffle operation penalizes any information about the temporal ordering that might leak into $\boldsymbol{f}$. Without this constraint, the model would have no incentive to separate static from dynamic information: $\boldsymbol{f}$ could absorb arbitrary content from the sequence, including information about its temporal dynamics, and the residual variables $\boldsymbol{\lambda}_{1:T}$ would carry only whatever $\boldsymbol{f}$ chose to ignore.

This observation points to something deeper. The shuffle constraint is the *only* mechanism in our framework that enforces the static/dynamic split. Its role is essentially that of an information bottleneck targeted at temporal content. Remove it, and the model is free to distribute informa-

tion between $\boldsymbol{f}$ and $\boldsymbol{\lambda}_{1:T}$ in whatever way is most convenient for reconstruction, which, under maximum-likelihood training, means that $\boldsymbol{f}$ will tend to capture as much of the sequence as possible, including its global dynamic structure.

Far from being a pathology, this behavior is actually desirable for a different purpose. If $\boldsymbol{f}$ can encode not only the static content of a video but also the overall structure of its motion (what is happening, how it unfolds, its global trajectory) then $\boldsymbol{f}$ becomes a *compact global descriptor* of the entire sequence. The residual variables $\boldsymbol{\lambda}_{1:T}$, for their part, need only capture the frame-specific deviations from this global reference, which are typically small and structured. The result is a representation architecture in which a single fixed-size vector summarizes an entire video, with temporal dynamics absorbed into the global latent rather than modeled frame-by-frame.

This is precisely the perspective explored in Chapter 5. The model architecture is identical to the one developed in the present chapter: a global encoder producing $\boldsymbol{f}$, combined with a conditional normalizing flow that maps per-frame latent codes through $\boldsymbol{f}$, but the shuffle constraint is entirely removed. The objective is no longer disentanglement but compact representation: we ask whether a single global latent variable, inferred with minimal temporal modeling, can serve as a general-purpose descriptor of video sequences.

The practical implications of such a representation are significant. A compact global video descriptor, obtained without expensive spatio temporal decoding, would be directly useful for downstream classification, retrieval, and clustering tasks, bypassing the need to handle temporal structure during fine-tuning. It would also enable a new paradigm for generative video modeling, in which generation reduces to synthesizing $\boldsymbol{f}$, an image-like object, followed by decoding individual frames as minimal residual variations around the global reference. For compression, conditioning all frames on a shared $\boldsymbol{f}$ reduces inter-frame redundancy in a principled way, potentially enabling more efficient video codecs.

> In summary, the two directions described here share a common intellectual lineage with the framework of the present chapter, but pursue complementary goals: one asks how to *disentangle more*, by extending functional independence from two factors to arbitrarily many and from sequential to static data; the other asks what happens when we *stop disentangling altogether* and let the global latent absorb as much in-

formation as the data allows. Together, they trace the boundaries of what the conditional normalizing flow representation can achieve and set the stage for Chapters 4 and 5.

4

Unsupervised Disentanglement Without Compromises: How Functional Orthogonality Enforces Identifiability

Adapted from:

Simon, M. C., Frossard, P., & De Vleeschouwer, C. (2026). Unsupervised disentanglement without compromises: How functional orthogonality enforces identifiability. In *Proceedings of the Forty-Third International Conference on Machine Learning (ICML 2026)*.

4.1 Introduction

In *Disentangled Representation Learning* (DRL), the main assumption is that data is generated from a small number of underlying low-dimensional latent factors, often referred to as concepts. The objective is to recover a representation that separates these concepts, ideally in an unsupervised manner [9, 155]. For instance, in facial images, underlying concepts might include hair length, hair color, or face orientation. But what exactly constitutes a concept? If unsupervised disentanglement is to be meaningful, it presupposes some form of intrinsic, human-free definition of the factors. Many prior works define concepts as statistically independent factors of variation [61, 84, 27] similarly to ICA [96, 113]. However, this assumption frequently fails to align with human intuition. Some semantically distinct concepts can be statistically dependent, while spurious correlations in the data may give rise to apparent dependencies between otherwise unrelated factors. To address these limitations, some approaches introduce causal modeling [167, 16, 100, 165, 150, 95], treating factors as nodes in a causal graph. Yet, enforcing causal assumptions requires strong priors and may obscure the practical goal of capturing semantically meaningful factors. Often, the aim is not to recover the true causal structure, but rather to identify and represent the factors and their relationships, especially in cases where the causal link is unclear or difficult to justify. For example, while hair length and gender are clearly distinct concepts, they often exhibit statistical dependence, yet a direct causal link between them is not well-defined.

Thus, we propose an alternative perspective on what constitutes a concept. We argue that, although concepts may exhibit statistical dependence, they should be understood as factors that influence observations through distinct and independent variations, not in the statistical sense, but in terms of their functional effects on the generative process. Each factor controls its own distinct aspect of the observation. This perspective aligns closely with the *Independent Causal Mechanisms* (ICM) principle [134] from causal inference [120, 123], which posits that the causal generative process of a system is composed of autonomous modules that do not influence each other. Applied outside a causal context, it can be formalized as an orthogonality condition on the columns of the Jacobian of the generative function mapping latent factors to observation [52]. This captures a form of local decoupling in how each factor contributes to the observed data. Unlike prior works

that treat this property as a sufficient condition to recover statistically independent sources [18, 52], we take a stronger stance: we argue that orthogonal influence is an *intrinsic defining property of meaningful concepts*. In contrast to these approaches, which rely on statistical independence and restricted functional families, we consider general nonlinear mixing functions with orthogonal Jacobians and explicitly allow for dependent latent factors. This perspective moves beyond the "Independent" in ICA and suggests that orthogonality, not independence, is the fundamental structural property underlying disentanglement.

Building on this framework, we show that *disentanglement emerges naturally from orthogonal influence*, without requiring statistical independence or supervision. This challenges the prevailing view that unsupervised disentanglement is fundamentally impossible, and demonstrates instead that meaningful disentanglement can arise *without compromise* provided that one adopts a functionally grounded notion of independence. We support our theoretical findings with empirical results and discuss implications for understanding why models such as Variational Autoencoders implicitly promote disentangled representations. More broadly, our work offers a principled step toward unifying disentanglement and causality under a common functional framework.

4.2 Related Works

Early approaches to disentangled representation learning were primarily inspired by *Independent Component Analysis* (ICA) [96, 113], aiming to recover statistically independent factors of variation from observed data in an unsupervised manner [88, 61]. However, as formalized by Locatello *et al.*[103], purely unsupervised disentanglement is fundamentally impossible without strong inductive biases. Specifically, there exist infinitely many equally valid latent representations that yield identical marginal distributions over observations, rendering the generative model non-identifiable. This impossibility statement mirrors classical findings in nonlinear ICA [79], where disentanglement has been proven to fail in the absence of additional assumptions or supervision [76, 75, 178, 78]. Consequently, subsequent works have sought to introduce suitable inductive biases, either through architectural constraints [14, 29, 104, 40, 141], weighting terms [27, 84], or partial supervision via auxiliary variables [105, 110, 55, 45, 77]. Comparatively, in this work, we focus on the fully unsupervised setting,

emphasizing the need for a more principled notion of factors that goes beyond mere statistical independence as used previously and that would lead to an identifiable representation.

Several lines of research have explored alternative formulations of disentanglement through group theory [179, 59], second-order Hessian-based penalties [122], Jacobian regularization [98] or directly encouraging orthogonality between latent directions [159, 22]. But, the first theoretical framework to formalize this idea of functional orthogonality was introduced by Gresele *et al.* [52], under the name *Independent Mechanism Analysis* (IMA). Subsequent works extended this framework to higher-dimensional latent spaces under the manifold hypothesis [49] and demonstrated that VAEs implicitly encourage IMA-like structure through their training objective [129, 5], providing a partial explanation for their empirical disentangling ability despite the absence of explicit supervision. The original IMA formulation showed that enforcing orthogonality in the Jacobian eliminates spurious and degenerate ICA solutions in specific nonlinear settings. Later studies generalized these findings, proving identifiability results for restricted classes of transformations such as conformal mappings [18, 178] or local isometry [67]. In contrast to these prior efforts, our work provides a comprehensive theoretical treatment of identifiability under the *general* case of orthogonal Jacobians, without restricting to specific functional families. Furthermore, whereas previous analyses assumed independent latent factors, we explicitly address the more realistic case of dependent concepts, thereby broadening the applicability of IMA to real-world generative processes and offering a unified framework for disentanglement beyond statistical independence.

4.3 Problem Formulation

We start by formalizing the data-generating process. We assume that observations $\boldsymbol{x} \in \mathbb{R}^n$ result from latent concepts $\boldsymbol{z} \in \mathbb{R}^d$, with $d \ll n$, that have undergone a smooth and invertible mapping $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^n$ which act as a mixing function

$$\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{z}), \qquad p(\boldsymbol{z}) = \prod_{i=1}^{d} p(\boldsymbol{z}_i | \boldsymbol{z}_{<i}), \tag{4.1}$$

where each $\boldsymbol{z}_i$, $i = 1, \ldots, d$, is an individual concept. The factorization is simply the chain rule of probability and holds for *any* ordering of the

components, expressing the joint distribution recursively without imposing any constraint on $p(\boldsymbol{z})$. Crucially, compared to *Blind Source Separation* (BSS) in [52], we do not assume the underlying concepts $\boldsymbol{z}$ to be independent and their distribution might be modeled recursively by a triangular mapping from independent noise, $z = \phi(\epsilon)$, e.g. with the Darmois construction [32, 79]. The goal is to learn an unmixing function $\hat{\boldsymbol{z}} = \boldsymbol{g}^{-1}(\boldsymbol{x})$ that would inverse $\boldsymbol{f}$ so that we recover the true concepts $\boldsymbol{z}$ up to some tolerable ambiguities. Formally, we adopt the definition of identifiability introduced in [52].

Definition 4.1. Let $\mathcal{F}$ be the set of all smooth and invertible functions $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^n$ and $\mathcal{P}$ be the set of all densities $p_{\boldsymbol{z}}$ with simply connected support on R^d. $\mathcal{M} \subseteq \mathcal{F} \times \mathcal{P}$ defines the subspace of models. The generative process is said to be *identifiable* on $\mathcal{M}$ if

$$\begin{aligned} &\forall (\boldsymbol{f}, p_{\boldsymbol{z}}), (\boldsymbol{g}, p_{\hat{\boldsymbol{z}}}) \in \mathcal{M}\colon\ \boldsymbol{f}_* p_{\boldsymbol{z}} = \boldsymbol{g}_* p_{\hat{\boldsymbol{z}}} \\ \Rightarrow\ &\exists \boldsymbol{P}, \boldsymbol{t} \text{ s.t. } (\boldsymbol{f}, p_{\boldsymbol{z}}) = (\boldsymbol{g} \circ \boldsymbol{P}^{-1} \circ \boldsymbol{t}^{-1}, (\boldsymbol{P} \circ \boldsymbol{t})_* p_{\hat{\boldsymbol{z}}}). \end{aligned} \tag{4.2}$$

Where $\boldsymbol{f}_* p_{\boldsymbol{z}}$ denotes the pushforward of $p_{\boldsymbol{z}}$ through $\boldsymbol{f}$, $\boldsymbol{P}$ is an arbitrary permutation, and $\boldsymbol{t}$ is an element-wise reparameterization of the concepts.

Identifiability [97] expresses the idea that, under suitable assumptions, a learned representation should coincide with the true generative factors up to well-characterized ambiguities: a permutation and a dimension-wise reparameterization. Such ambiguities are acceptable as they preserve the disentangled structure of the representation. Without canonical choice of coordinate system, different parameterizations remain equally valid.

Unfortunately, such identifiability cannot be achieved under the current formulation without further constraints. To illustrate this, consider the case where the latent variables $\boldsymbol{z}$ are generated from independent noise via an arbitrary function ϕ, and subsequently mapped to observations via another bijection $\boldsymbol{f}$. In the absence of additional assumptions, any pair $(\phi, \boldsymbol{f})$ yielding the same distribution over $\boldsymbol{x}$ is equally valid. This means that $\boldsymbol{z}$ could be any arbitrary entangled transformation of the true underlying factors, with dependencies entirely absorbed by ϕ. Consequently, the same observed distribution can be explained by many distinct and incompatible latent structures. The model is thus underdetermined. This lack of identifiability undermines the interpretability and meaning of the latent factors: if the ground-truth generative process is not uniquely recoverable,

then the notion of concepts as fundamental explanatory variables loses its significance. To resolve this ambiguity, additional conditions are necessary to restrict the solution space and render the problem well-posed. We argue, following prior works [18, 52], that what is missing is a notion of independence, not in the statistical sense, but in terms of the functional influence that each factor exerts on the observations. Mathematically, we can formalize this as

Assumption 4.1. Let $\boldsymbol{J}_f$ denote the Jacobian of the generative function $\boldsymbol{f}$, and let $\partial \boldsymbol{f}/\partial z_i$ denote its i-th column. Each concept z_i affects the observations through a distinct mode of variation:

$$\forall \boldsymbol{f} \in \mathcal{F}, \quad \forall i \neq j: \quad \frac{\partial \boldsymbol{f}}{\partial z_i}^T \frac{\partial \boldsymbol{f}}{\partial z_j} = 0, \tag{4.3}$$

i.e., the partial derivatives of $\boldsymbol{f}$ with respect to distinct latent variables are orthogonal at every point. Equivalently, $\boldsymbol{J}_f^T \boldsymbol{J}_f$ is diagonal everywhere.

This orthogonality condition enforces that perturbations along different latent dimensions induce locally independent variations in the data space. It provides a geometric form of disentanglement by requiring that the effect of each latent factor on the observations be decoupled from the others. Importantly, this assumption applies to the entire model class $\mathcal{M}$ and is posited as a defining property of meaningful factors. In doing so, it shifts the inductive bias from statistical independence of the latent variables to independence of their functional contributions, which we argue is more aligned with the way humans intuitively define and distinguish concepts. Among the functions satisfying Asm.4.1, a notable subclass is formed by conformal maps, characterized by Jacobians of the form $\boldsymbol{J}_f = \lambda(\boldsymbol{z})\boldsymbol{Q}(\boldsymbol{z})$, where $\lambda : \mathbb{R}^d \to \mathbb{R}^+$ is a smooth scalar field and $\boldsymbol{Q}(\boldsymbol{z}) \in O(d)$ is an orthogonal matrix. Conformal maps preserve local angles. A visualization is provided in Fig.4.1.

Another way to understand this assumption is through the lens of generative invertibility. When the latent variables $\boldsymbol{z}$ are constructed recursively via a function ϕ from independent noise $\boldsymbol{\epsilon}$, the Jacobian of ϕ becomes lower triangular due to the autoregressive structure of the dependencies. In this setting, the full generative mapping from noise to observations is $\boldsymbol{x} = \boldsymbol{f} \circ \phi(\boldsymbol{\epsilon})$, and its Jacobian is the product $\boldsymbol{J}_f \cdot \boldsymbol{J}_\phi$. If $\boldsymbol{J}_\phi$ is lower triangular and $\boldsymbol{J}_f$ has orthogonal columns, then this product is reminiscent of a QR

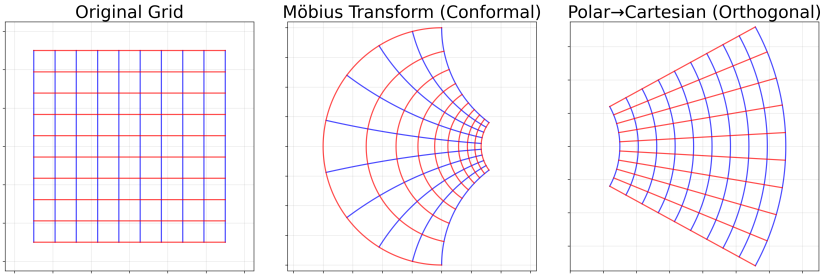


Fig. 4.1 Grid Transformations: Visualizing Orthogonal Jacobians.

decomposition of an invertible matrix. Since the QR decomposition of an invertible matrix is unique up to signs [50], this analogy hints at a form of uniqueness in the factorization of the generative process. By encouraging orthogonality in f, we effectively constrain the overall generative model in a way that helps disambiguate the latent representation. This supports the idea that the orthogonality of functional contributions is not merely a heuristic, but a theoretically grounded principle that facilitates identifiability in the absence of statistical independence.

4.4 Identifiability Conditions

We now turn to the central theoretical contribution of this work: establishing that the orthogonality constraint on the Jacobian of the generative function yields a uniquely recoverable generative model, up to the usual ambiguities. This result highlights that orthogonality is not merely a convenient regularization, but rather a *fundamental property that defines what constitutes a factor in the generative process*. Without identifiability, multiple entangled latent representations can explain the same observed distribution, rendering the notion of "factors" devoid of semantic meaning. Demonstrating identifiability therefore establishes the conceptual and mathematical significance of orthogonal functional influence as the defining criterion for meaningful disentanglement. In the following, we analyze identifiability in two stages. We first consider the case where the latent factors are statistically independent. We then relax this assumption and extend the analysis to the more general and realistic case where the latent factors may exhibit statistical dependencies.

4.4.1 Identifiability Under the Independence Assumption

We first consider the case where the latent factors are statistically independent. This setting is closely related to classical BSS and ICA, where the goal is to recover the independent sources that generate the observed data. Here, we extend this principle to the nonlinear setting under the orthogonality constraint introduced in Asm.4.1.

Proposition 4.1. *Let the observed data $\boldsymbol{x} \in \mathbb{R}^n$ lie on a d-dimensional manifold generated by latent variables $\boldsymbol{z} \in \mathbb{R}^d$, with $d \ll n$, through a smooth and invertible mapping $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^n$ satisfying Asm.4.1. Each latent coordinate z_i corresponds to a distinct underlying concept. Suppose further that the latent variables are statistically independent, $p(\boldsymbol{z}) = \prod_i p(\boldsymbol{z}_i)$, and that at most one component z_i follows a Gaussian distribution. Then, the generative process $(\boldsymbol{f}, p_{\boldsymbol{z}})$ is identifiable in the sense of Def.4.1.*

The complete proof is provided in App.B.1. Intuitively, we can illustrate it in the special case where $\boldsymbol{f}$ is conformal and show that identifiability follows directly. Indeed, using the same notations as in Def.4.1, denote $\boldsymbol{h} = \boldsymbol{f}^{-1} \circ \boldsymbol{g}$ the function mapping $\hat{\boldsymbol{z}}$ to $\boldsymbol{z}$. Independence implies $\det \boldsymbol{J}\boldsymbol{h} = 1$ up to a coordinate-wise reparameterization. This, combined with the conformal structure of $\boldsymbol{h}$ (composition of conformal mappings), forces $\boldsymbol{h}$ to be both angle and volume preserving. Hence, $\boldsymbol{h}$ reduces to a rigid transformation (rotation and translation), yielding the same identifiability condition as in linear ICA. This provides a simple and direct proof of identifiability in the conformal case compared to [18] or [178].

More generally, Prop.4.1 shows that the orthogonality constraint ensures identifiability even beyond the conformal case for the general orthogonal Jacobian. However, that simple proof reveals a conceptual link between our framework and classical ICA : both aim to recover latent directions that are statistically independent and Asm.4.1 naturally collapses the nonlinear problem to the identifiable linear ICA regime, recovering the condition that identifiability holds when at most one source is Gaussian. This is consistent with prior results in [52] for restricted functional classes and explains why the same conclusions as ICA apply. In this sense, Asm.4.1 generalizes the linear orthogonality condition underlying ICA to the nonlinear regime. This connection also relates to Principal Component Analysis (PCA) [1], where orthogonality plays a central role, the orthogonal Jacobian assumption can be interpreted as a local, nonlinear analogue

of PCA, while statistical independence extends it globally to ICA.

Overall, this result goes beyond SoTA, as it directly challenges the widely accepted impossibility result which asserts that unsupervised disentanglement is unachievable [79, 103]. Our findings reveal that, when the generative process satisfies a functional independence condition, the generative process becomes identifiable even in the absence of supervision. This helps explain why methods such as VAEs and other architectures implicitly or explicitly enforcing local orthogonality often succeed empirically in disentangling latent factors, particularly on synthetic or structured datasets that directly satisfy the assumptions.

Finally, Prop.4.1 emphasizes the dual nature of the identifiability problem: (1) a **local** aspect, governed by the Jacobian orthogonality that determines how latent factors influence the observations; and (2) a **global** aspect, governed by the distribution $p(\boldsymbol{z})$ that constrains how these factors are statistically organized. Together, these complementary constraints define a principled pathway to achieving meaningful and identifiable disentanglement.

4.4.2 Identifiability Under Dependent Factors

In practice, and as discussed in the introduction, the assumption of statistically independent latent factors rarely holds, either due to inherent relationships or spurious correlations in the dataset. It is therefore essential to study identifiability when factors are potentially dependent. In this setting, the data $\boldsymbol{x}$ is still generated by a smooth and invertible mapping $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^n$ satisfying Asm.4.1, but the latent variables $\boldsymbol{z}$ no longer follow a factorized distribution. Instead, using the probability chain rule, their joint distribution can be modeled recursively as $p(\boldsymbol{z}) = \prod_{i=1}^{d} p(z_i \mid z_{<i})$ which can equivalently be expressed through a triangular transformation $\phi : \mathbb{R}^d \to \mathbb{R}^d$ of independent noise variables $\boldsymbol{\epsilon}$ [32], i.e., $\boldsymbol{z} = \phi(\boldsymbol{\epsilon})$. Importantly, we do not impose any causal interpretation on ϕ; it merely serves as a generic mechanism to model dependent factors.

However, in this more general case, identifiability is no longer guaranteed. As established in Prop.4.1, identifiability arises from the interaction of two complementary constraints: a *local geometric constraint* and a *global statistical constraint*. When independence is removed, the global constraint disappears, leaving the model underdetermined. Indeed, a trivial counterexample illustrates this failure: consider $\boldsymbol{f}$ as the identity map-

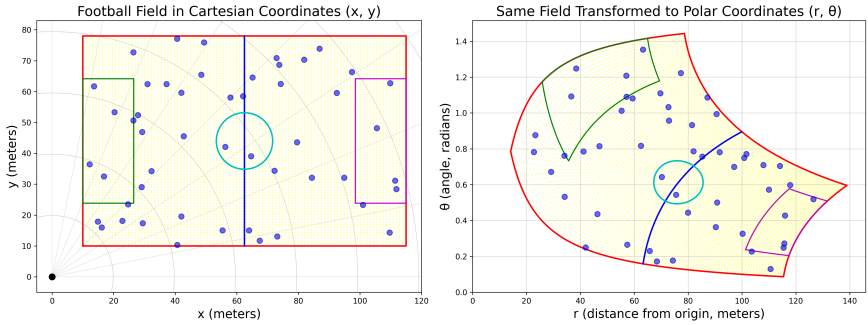


Fig. 4.2 Assume we aim to represent the position of a ball on a field as a two-dimensional latent variable. The most direct representation uses Cartesian coordinates (x, y) (left), regardless of the distribution of positions. However, an equally valid alternative could be a polar representation (r, θ) (right). Without additional assumptions, there is no principled way to identify which representation is the "true" one: (r, θ) is simply an entangled reparameterization of (x, y). Under an independence assumption, however, a privileged representation emerges; for example, if positions are uniformly distributed over the field, this constraint favors the Cartesian coordinates (x, y). In the absence of such assumptions, these entangled representations are indistinguishable. Nevertheless, we argue that even without independence, certain representations are preferable. Specifically, those in which all combinations of factor values can in principle occur, as with the (x, y) coordinates in this example (Asm.4.2).

ping, which trivially satisfies the orthogonality condition. In this case, $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{z}) = \boldsymbol{z}$, and the overall generative model $\boldsymbol{x} = \boldsymbol{f} \circ \boldsymbol{\phi}(\boldsymbol{\epsilon})$ can represent any data distribution $p(\boldsymbol{x})$ by appropriately choosing $\boldsymbol{\phi}$ which can represent any distribution e.g., with the Darmois construction [32]. Thus, the model perfectly explains the data without uncovering any meaningful latent structure. This demonstrates that, in the absence of additional constraints, the concept of latent factors becomes vacuous: if multiple equally valid generative decompositions exist, then the notion of disentanglement loses its interpretability. An illustration of this degeneracy is provided in Fig.4.2.

A natural way to restore identifiability would be to incorporate auxiliary variables or weak supervision, such as labels or side information, to anchor the latent space and help align the learned axes with meaningful factors [83]. However, such information is not always available and introduces an undesirable degree of supervision.

These limitations motivate the search for a weaker, yet sufficiently informative, *global* constraint that can complement the *local* orthogonality condition. The key intuition is that orthogonality of the Jacobian already enforces strong geometric rigidity. Consequently, even limited global information about the latent domain may suffice to fully determine the mapping. In the theory of conformal mappings for dimensions $d > 2$, Liouville's theorem [13] implies that any conformal transformation of $\mathbb{R}^d$ must be a Möbius transformation, i.e., a composition of translations, rotations, dilations, and inversions. When the data lies on a bounded, simply connected domain $\Omega \subset \mathbb{R}^d$ that is not invariant under inversion (a property satisfied by a broad class of supports, including all polytopes and most bounded smooth domains) non-rigid Möbius transformations are ruled out. This implies that, *when the latent support is known, the generative process becomes identifiable, up to the natural rigid symmetries of that support*. Formal statements and illustrative examples are provided in App.B.2, where we also discuss the special case $d = 2$ via the Riemann Mapping Theorem.

Of course, such detailed geometric knowledge of the latent domain is rarely known a priori. Nevertheless, the above discussion suggests that identifiability can be recovered by introducing a structural assumption on the latent distribution that serves a similar purpose, thereby providing the required global constraint.

Assumption 4.2. The latent variables $\mathbf{z} \in \mathbb{R}^d$ follow a distribution $p_{\mathbf{z}}$ supported on a bounded, simply connected region $\Omega \subset \mathbb{R}^d$ such that, up to a continuous strictly monotone coordinate-wise reparameterization, Ω is a hypercube. Equivalently, the copula associated with $p_{\mathbf{z}}$ has full support on $[0,1]^d$.

Intuitively, Asm.4.2 states that all combinations of latent factors are possible, that is, each latent coordinate can vary within its range, regardless of the values taken by the others. One may think of the latent dimensions as concept "sliders" that can be adjusted to produce every possible combination of factors. This implies that each factor corresponds to a distinct, manipulable degree of freedom. For example, consider the concepts of altitude and temperature. Although these quantities are correlated in nature, it is perfectly possible to observe configurations that break this dependency (e.g., a heated room at high altitude). The ability to realize or observe all such combinations is precisely what gives meaning to the notion of sep-

arate factors. Hence, Asm.4.2 can be viewed as a non-causal analog of interventions in causal models, ensuring that the dataset is rich enough to cover the joint range of the latent variables. A detailed discussion of both assumptions is provided in App.B.1.4.

Interestingly, if the factors are statistically independent and supported on a bounded domain, Asm.4.2 is fulfilled and thus this assumption is a direct generalization of the case of independent factors for bounded latent spaces. Critically, without such a global constraint, many different valid representations may satisfy the local functional independence enforced by an orthogonal Jacobian. Among these representations, however, there typically exists a *preferential* one that aligns with the combinatorial structure of the latent domain; this phenomenon is illustrated in Fig.4.2. Using this conclusion, we obtain the following result.

Proposition 4.2. *Let the observed data $\boldsymbol{x} \in \mathbb{R}^n$ lie on a d-dimensional manifold generated by latent variables $\boldsymbol{z} \in \mathbb{R}^d$, with $d \ll n$, through a smooth and invertible mapping $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^n$ satisfying Asm.4.1. Suppose further that the latent distribution $p_{\boldsymbol{z}}$ satisfies Asm.4.2. Then, the generative process $(\boldsymbol{f}, p_{\boldsymbol{z}})$ is identifiable in the sense of Def.4.1.*

The proof is given in App.B.1. Together with the local orthogonality condition, Asm.4.2 ensures a unique alignment. This result can be interpreted as a general principle:

> *Identifiability, and hence the very notion of a "concept" or "factor", emerges from the dual constraints that define it:*
>
> 1. **Local constraint (Assumption 4.1):** factors influence the data through orthogonal directions, capturing functional independence in the generative mechanism;
> 2. **Global constraint (Assumption 4.2):** the latent domain allows all possible factor combinations; each factor corresponds to a distinct manipulable degree of freedom.
>
> Once these defining properties of latent factors are in place, identifiability follows naturally. This observation forms the central insight of our theoretical framework: *disentanglement is not an artifact of independence or causality, but a structural property emerging from the interplay between local orthogonality and global combinatorial completeness.*

We next turn to the empirical validation of these results.

4.5 Experiments

Theoretical Prop.4.1 and 4.2 establish that, under orthogonal functional influence, the generative model becomes identifiable. In this section, we empirically validate these results by investigating whether models trained solely on observations $\boldsymbol{x}$ can recover the underlying latent concepts $\boldsymbol{z}$ up to the admissible ambiguities specified in Def.4.1. Concretely, we compare models trained with and without an orthogonality constraint on the Jacobian of the generative mapping and evaluate their ability to disentangle latent factors. Details on the experiments are provided in App.B.3.

Data Generation. We consider two regimes for the latent variables $\boldsymbol{z} \in \mathbb{R}^d$, with $d \in \{3, 6, 9\}$: (i) Independent factors, latent variables are sampled independently from a uniform distribution on $[0, 1]^d$ (ii) Dependent factors, latent variables are generated using a flexible *Normalizing Flow* (NF) [91], that induces strong statistical dependencies while ensuring the resulting copula has full support on $[0, 1]^d$. This construction explicitly violates independence while satisfying Asm.4.2. In both cases, latent variables are mapped to observations via a smooth and invertible ground-truth mixing $\boldsymbol{f}$. We consider two classes of generative mappings: (i) conformal Möbius transformations; and (ii) more general nonlinear mappings with orthogonal Jacobians of the form $\boldsymbol{J}_f = \boldsymbol{Q}(\boldsymbol{z})\boldsymbol{D}(\boldsymbol{z})$, obtained by composing conformal and non-conformal transformations. This second class goes strictly beyond conformal mappings and tests identifiability under general orthogonal Jacobians.

Models. To model the data distribution $p(\boldsymbol{x})$, we train expressive Residual Flows [26] with full Jacobians. We learn both *Unconstrained baselines* with no structural constraints on the Jacobian and *Orthogonally constrained models* using the C_{IMA} loss introduced by Gresele *et al.* [52] to regularize the maximum likelihood objective, enforcing Asm.4.1. In the independent setting, the base distribution is a standard Gaussian with diagonal covariance. In the dependent setting, the base distribution is itself learned using a RealNVP flow [36] with a triangular Jacobian, ensuring a full-support copula while modeling dependencies between latent variables.

Evaluation Metrics. For evaluation, we compute (i) the KL divergence between the learned model and the true data-generating distribution to assess density estimation quality; (ii) the *Mean Correlation Coefficient* (MCC)

between the recovered and ground-truth latent variables [83]; and (iii) a nonlinear extension of the Amari distance [52] between the true mixing and the learned unmixing. The latter equals zero if and only if the model is identifiable in the sense of Def.4.1. As all models managed to fit the data and have equivalent KL divergence metrics we only report the MCC and Amari distance to measure disentanglement.

Results. Fig.4.3 reports MCC and Amari distances for unconstrained and orthogonally constrained models, across independent and dependent latent variables, for both conformal and general $\boldsymbol{QD}$ generative mappings. For each configuration, 30 independent runs are performed, yielding a total of 720 trained models. As expected, standard NFs consistently fail to recover the latent factors despite achieving good likelihoods, confirming that the problem remains underdetermined. In contrast, introducing the orthogonality constraint leads to substantial improvements in both MCC and Amari distance, often approaching perfect identifiability. These results empirically validate Prop.4.1 and 4.2. Notably, while prior work largely focused on conformal mappings, our results demonstrate that identifiability extends to significantly more general nonlinear transformations with orthogonal Jacobians. Strikingly, performance is also comparable in the independent and dependent settings. This provides strong empirical evidence that disentanglement does not fundamentally rely on statistical independence or causal assumptions. Instead, meaningful and identifiable representations of dependent concepts can be recovered purely from orthogonal functional influence, supporting the central thesis of this work. Additional results and qualitative visualizations are provided in App.B.4, including comparison with VAEs across standard disentanglement metrics.

Finally, Fig.4.4 presents some qualitative results illustrating the role of global domain information. We generate data with varying latent supports and train a Neural Spline Flow [39] equipped with orthogonal Jacobian constraints. Consistent with the discussion in Sec.4.4.2, the learned representations are identifiable up to the rigid symmetries of the latent domain when its structure is known. This behavior aligns with the theoretical analysis in App.B.2 and further illustrates the interplay between local orthogonality and global constraints.

Overall, these experiments demonstrate that enforcing orthogonal functional influence is sufficient to enable unsupervised disentanglement. The empirical findings closely match the theoretical predictions, providing strong support for the proposed notion of concepts.

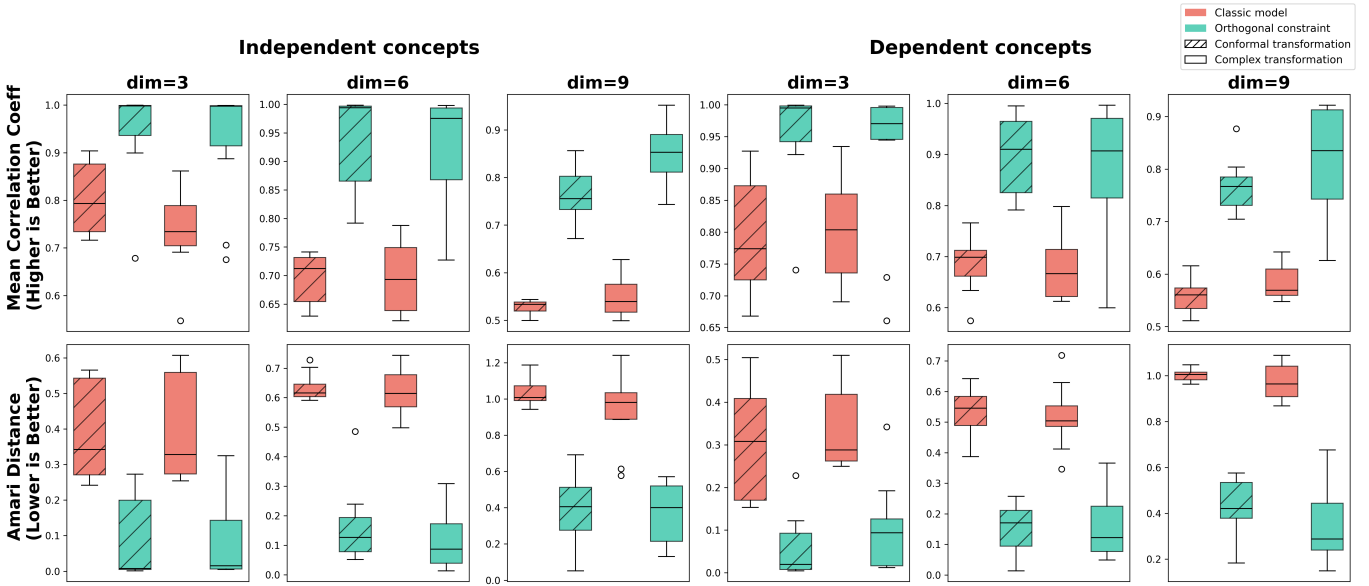


Fig. 4.3 MCC and Amari distances quantitative results for $d \in \{3, 6, 9\}$ in the independent/dependent cases.

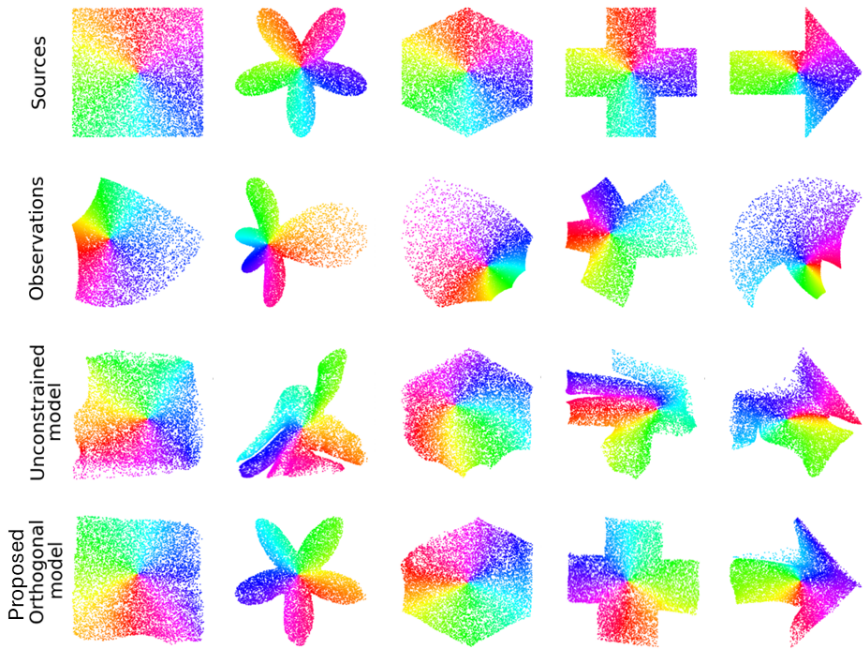


Fig. 4.4 Qualitative results under various sources domain.

4.6 Discussion and Perspectives

We now discuss some broader implications of our findings and outline several promising directions for future work.

Causal Representation. Throughout the chapter, we have drawn connections to causal representation learning, whose aim is to recover latent factors together with their causal structure. However, in many applications, enforcing a full causal model is often unrealistic: the causal relations between factors are unclear or not identifiable from observational data alone. Our results show that meaningful factors can nevertheless be recovered by leveraging causally inspired principles, such as ICM, without committing to an explicit causal graph. This suggests an alternative perspective on causal representation learning, in which causal assumptions are relaxed to prioritize factor recovery. If causal structure is ultimately required, it can

be inferred *after* disentanglement, using additional assumptions, domain knowledge, or limited interventions, effectively deferring causal resolution to a second stage. This "best effort with the available data" philosophy aligns with the practical reality of many machine-learning applications, where full causal supervision is rarely accessible. Overall, our results emphasize that it is neither statistical independence nor causal directionality that fundamentally defines a latent factor.

Importantly, our generative model remains compatible with causal analysis when such structure is desired. In our framework, the observation map f has an orthogonal Jacobian, while dependencies between latent variables are modeled via a triangular transformation ϕ. In this setting, results by Xi *et al.* [162] imply that such models recover representations up to the Markov Equivalence Class. Consequently, our approach can be viewed as a strong starting point for causal discovery, with causal ordering potentially resolvable using only limited interventional information.

VAE Disentanglement. Finally, we revisit *Variational Autoencoders* (VAEs) through the lens of our framework. VAEs are among the most widely used models for unsupervised disentanglement. A VAE learns a latent variable model by jointly optimizing a reconstruction term, $-\log p(x \mid z)$, and a regularization term enforcing the approximate posterior $q(z \mid x)$ to match a factorized Gaussian prior $p(z) = \mathcal{N}(0, I)$. A well-established belief is that the factorized prior is the primary mechanism responsible for disentanglement. Works such as [27, 84] formalize this viewpoint by decomposing the evidence lower bound (ELBO) as

$$\mathcal{L}_{\text{VAE}} = -\log p(x \mid z) + \mathrm{KL}(q(z \mid x) \,\|\, p(z)) \tag{4.4}$$

$$= \underbrace{-\log p(x \mid z)}_{\text{Reconstruction}} + \underbrace{\mathrm{KL}\Big(q(z) \,\|\, \textstyle\prod_j q(z_j)\Big)}_{\text{Total Correlation}} + \underbrace{\textstyle\sum_j \mathrm{KL}(q(z_j) \,\|\, p(z_j))}_{\text{Dimension-wise KL}} + \underbrace{I(x;z)}_{\text{Mutual Information}} . \tag{4.5}$$

The dimension-wise KL encourages each latent dimension to match a univariate Gaussian, hence simply affecting the reparameterization, and the mutual-information term is undesirable. In this decomposition, the total correlation term penalizes statistical dependence between latent coordinates and is therefore viewed as the component driving disentanglement. However, this explanation is incomplete. Indeed, we show here below that

models such as Normalizing Flows share exactly the same total-correlation penalty while they do **not** generally produce disentangled representations.

Consider a NF, which learns an invertible mapping $\boldsymbol{h}$ s.t.

$$\log p(x) = \log p(z) + \log\left|\det(\mathrm{d}h(x)/\mathrm{d}x)\right|, \tag{4.6}$$

with $p(z) = \mathcal{N}(0, I)$. Since the pushforward distribution $q(z)$ induced by h does not necessarily match $p(z)$, training effectively minimizes cross-entropy:

$$\begin{aligned}\mathcal{L}_{\mathrm{NF}} &= -\log q(z) + \mathrm{KL}(q(z) \parallel p(z)) - \log\left|\det(J)\right| &(4.7)\\ &= -\log p(x) + \underbrace{\mathrm{KL}\Big(q(z) \parallel \textstyle\prod_j q(z_j)\Big)}_{\text{Total Correlation}} \\ &\quad + \underbrace{\textstyle\sum_j \mathrm{KL}(q(z_j) \parallel p(z_j))}_{\text{Dimension-wise KL}}. &(4.8)\end{aligned}$$

Thus, flows exhibit the same global independence pressure as VAEs. Yet, in practice, flows remain non-disentangled in their standard form [36, 175]. This discrepancy highlights that factorized priors, or equivalently, total-correlation minimization, are insufficient to explain VAE's natural bias toward disentanglement.

The crucial missing ingredient is the *factorized approximate posterior* which is often overlooked. VAEs simultaneously enforce:

1. a factorized prior $p(z)$, promoting independence across data points (global structure), and
2. a diagonal-covariance approximate posterior $q(z \mid x) = \mathcal{N}(\mu, \sigma^2 I)$, promoting locally axis-aligned variations within each data point, hence orthogonality.

The fact that the factorized posterior implicitly induces an orthogonal Jacobian of the encoder has been formally proven in [129, 5]. Consequently, disentanglement in VAEs arises not merely from independence constraints but from the *joint* imposition of (i) global independence via the prior and (ii) local orthogonality via the posterior. This dual constraint aligns precisely with the structure promoted by our framework. It also explains why VAEs succeed on datasets with well-separated generative factors, typically synthetic data, and fail on more complex settings, e.g., with dependent structure.

To further validate this dual-mechanism explanation, we design additional experiments on standard image disentanglement benchmarks, namely *dSprites* [109] and *3DShapes* [19]. In Sec.4.5, we considered models without orthogonality bias (NFs) and showed that adding this bias improves disentanglement. Here, we follow the complementary approach: we start from a model that includes both independence and orthogonality (a VAE), and we remove the orthogonality bias to evaluate its effect.

The mechanism by which VAEs implicitly enforce Jacobian orthogonality has been analyzed in prior work [129]. In simplified terms, the factorized Gaussian posterior enforces a row-orthogonal encoder Jacobian, and in the near-deterministic regime the decoder approximately inverts the encoder, resulting in a column-orthogonal decoder Jacobian. Based on this observation, we construct a variant of the β-VAE where we replace the diagonal Gaussian posterior with a NF posterior. This is a standard technique in the VAE literature to increase expressiveness and better match the prior, thereby improving density modeling [90]. However, unlike the factorized Gaussian case, this more expressive posterior no longer enforces a diagonal covariance structure, and thus removes the implicit Jacobian orthogonality constraint. Importantly, this change leaves the overall objective unchanged, particularly the total-correlation/independence pressure imposed by the prior, thereby isolating the effect of removing the orthogonality bias while preserving the independence mechanism.

We implement this model following [103] and reproduce standard β-VAE results. We then train the exact same model, with identical architecture and training procedure, but using a flow-based posterior (denoted β-flowVAE). The results are reported in Tab.4.1. If orthogonality were not important, this more expressive model should yield improved disentanglement, since it better matches the prior hence improving independence. However, we observe a systematic substantial decrease in disentanglement performance. The key difference is precisely the removal of the orthogonality bias. This provides direct empirical evidence that independence alone is not sufficient, and that orthogonality plays a crucial role, further confirming our claims on the sources of disentanglement.

Beyond VAE's. Although VAEs remain the canonical model for disentanglement, they suffer from the well-documented *information preference* or *posterior collapse* phenomenon [29, 176]. Intuitively, due to the explicit minimization of the mutual information $I(x;z)$ in Eq.4.4, VAEs trade off expressive reconstruction against informative latent representations, often result-

Table 4.1 Disentanglement performance averaged over 10 random seeds across datasets and models.

		Standard VAE models W/ Ortho. J.			W/O Ortho. J.
Dataset	Metric	β-VAE	FactorVAE	β-TCVAE	β-FlowVAE
dSprites	DCI dis.	0.1879 ± 0.0967	0.2527 ± 0.0617	0.2963 ± 0.0905	0.0975 ± 0.0307
	MIG	0.1070 ± 0.0706	0.1643 ± 0.0654	0.1783 ± 0.0689	0.0443 ± 0.0194
	SAP	0.0465 ± 0.0338	0.0673 ± 0.0119	0.0663 ± 0.0141	0.0195 ± 0.0062
	β-VAE score	0.8238 ± 0.0758	0.8548 ± 0.0225	0.8628 ± 0.0345	0.7329 ± 0.0665
	FactorVAE score	0.6523 ± 0.1018	0.7354 ± 0.0973	0.7312 ± 0.1204	0.6020 ± 0.0997
3DShapes	DCI dis.	0.5349 ± 0.3034	0.6762 ± 0.0847	0.6539 ± 0.3033	0.1844 ± 0.0386
	MIG	0.2279 ± 0.1791	0.2977 ± 0.1344	0.2899 ± 0.2412	0.0617 ± 0.0215
	SAP	0.0662 ± 0.0460	0.0707 ± 0.0320	0.0811 ± 0.0490	0.0265 ± 0.0112
	β-VAE score	0.9675 ± 0.0621	0.9477 ± 0.0741	0.9756 ± 0.0520	0.8031 ± 0.0596
	FactorVAE score	0.8386 ± 0.1014	0.8234 ± 0.0676	0.8725 ± 0.1403	0.7142 ± 0.0834

ing in blurry reconstructions, weak generative fidelity, and latent variables that carry minimal semantic signal.

Our analysis shows that disentanglement relies on independence and orthogonality constraints, not on the VAE objective per se. This opens the door to transplanting these inductive biases into alternative generative models that avoid VAE-specific limitations. In Fig.4.3, we demonstrated that NFs, which do not naturally produce disentangled representations, can be endowed with disentanglement capabilities simply by augmenting them with an orthogonal regularization. Unlike VAEs, flows optimize exact likelihoods and do not include a mutual information penalty in Eq.4.7, thereby avoiding the same reconstruction–information trade-off.

This perspective reinforces the relevance of recent approaches such as orthogonality regularized GANs [159] or diffusion models [28, 124]. Our framework provides a theoretical justification for why these methods succeed. We argue that orthogonality should be viewed as a fundamental design principle for concept-structured and disentangled representations. By clarifying the mechanisms through which disentanglement arises, our framework illuminates how to systematically design new models that embody these principles. Overall, these insights indicate a promising direction for future research aimed at developing principled generative models with inherently disentangled representations. In particular, developing scalable methods to enforce or approximate Jacobian orthogonality efficiently in high-dimensional settings remains an important open problem, and a natural next step toward making the theoretical guarantees established in this work practically applicable at scale.

4.7 Conclusion

We showed that disentanglement can be achieved without statistical independence or causal supervision by defining concepts through orthogonal functional influence. Under this notion, nonlinear generative models with orthogonal Jacobians are identifiable up to standard ambiguities, even with dependent latent factors, extending prior identifiability results to the general orthogonal case. This framework explains the empirical disentanglement observed in models such as VAEs, clarifies why likelihood-based models fail without geometric constraints, and establishes orthogonality as a fundamental principle for identifiable and meaningful representation learning.

5

All for One and One for All: Learning a Global Latent Representation for Video Sequences

Adapted from:

Simon, M. C., Frossard, P., & De Vleeschouwer, C. (2026). All for one and one for all: Learning a global latent representation for video sequences. *Manuscript currently under review*.

5.1 Introduction

Learning minimal sufficient summaries of complex data distributions is a central challenge in probabilistic modeling. For static image data, this problem has largely been addressed through Autoencoders (AEs) [62] where a single compact latent code summarizes an image, enabling downstream tasks. For sequential data and video, however, no such standard compact representation exists. A natural extension of the image AE assigns a latent code to each frame of the sequence. This retains reconstruction fidelity but loses the compactness that makes image representations so useful: the representation is now a *sequence* of codes rather than a single element [101, 164, 166]. As a result, downstream tasks must handle temporal structure explicitly, and models trained on such representations must enforce temporal coherence, a non-trivial requirement that often demands expensive temporal architectures.

Representing an entire video with a *single* compact latent would be highly desirable, yet remains largely unexplored. In this work, we thus raise the question whether a video sequence can be summarized in such a way. A direct extension of image-based approaches would require a decoder capable of reconstructing a full sequence from a single latent, which inherently demands modeling temporal dependencies. One could treat time as an additional spatial dimension, leading to costly spatio-temporal architectures [63, 66, 160], or rely on autoregressive or attention-based temporal modeling [136, 145, 73]. In both cases, the burden of capturing temporal structure is placed on the decoder, and compactness is traded off against modeling complexity.

We propose a different perspective. Rather than increasing the capacity of temporal models, we investigate whether temporal structure can be absorbed into the global representation itself. Specifically, we hypothesize that, for many video sequences, a single global latent element can encode not only the static content (e.g., scene layout, object identities, background) but also the global structure of the motion: what is happening, how it unfolds, its overall trajectory. The frame-to-frame variability that remains, once this global structure is accounted for, is minimal and structured. We therefore introduce a global latent variable $\boldsymbol{s} \in \mathcal{S}$, inferred from the full sequence $\boldsymbol{x}_{1:T}$, which summarizes the entire video into a single image-like element serving as a learned frame of reference. Each frame $\boldsymbol{x}_t$ is then

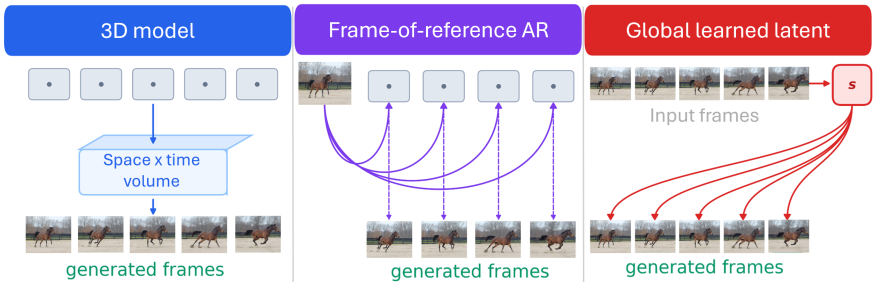


Fig. 5.1 Strategies for sequence modeling. Left: joint spatio-temporal modeling (expensive). Middle: frame-wise with sequential modeling. Right: global latent $\boldsymbol{s}$ acting as a learned frame of reference with residual temporal dynamics.

modeled conditionally on $\boldsymbol{s}$, such that the temporal structure of the sequence is captured through the relation of each frame to this shared global representation rather than through explicit sequential dependencies. This formulation provides a third alternative to spatio-temporal expansion and autoregressive decoding, as illustrated in Fig.5.1.

This factorization shifts the role of representation learning: instead of distributing information across time, the model is encouraged to concentrate it into a single latent variable of fixed size. Crucially, because $\boldsymbol{s}$ lies in an image-like space, the resulting representation can be used directly for downstream tasks without requiring temporal reasoning. Our framework builds upon the theoretical and practical properties of Variational AutoEncoders (VAEs) [88] and Normalizing Flows (NFs) [130, 89].The global latent variable $\boldsymbol{s}$ is optimized to maximize the likelihood of the observed frames, thereby encouraging it to encode maximal information about the video. This perspective reframes video representation learning as the problem of learning a minimal, sufficient, and fixed-size summary of a sequence in a fully unsupervised manner.

In the remainder of the chapter, we formalize the proposed model, analyze its properties, and discuss its implications for **generative modeling**, **compression**, and **representation learning**. We further provide empirical evidence demonstrating that a single learned global element can effectively encode the essential structure of video sequences without requiring any direct temporal modeling.

5.2 Related Works

A large body of work has explored global latent representations for video by factorizing sequences into static content and dynamic motion [146, 69, 99, 7, 141]. In these approaches, the global latent is explicitly constrained to capture time-invariant information, while temporal dynamics are modeled through additional frame-wise or sequential variables. While effective for disentanglement, this construction inherently limits the expressivity of the global representation, preventing it from capturing the full structure of the sequence. In contrast, we pursue the opposite objective: rather than separating content and motion, we investigate whether a video sequence can be summarized compactly into a single global latent variable, concentrating both content and dynamics into a unified representation, and restricting local variables to modeling only residual variation. A related conceptual direction is given by Neural Processes [172, 47, 48, 85], which compress a set of observations into a single latent representation for prediction; however, their objective is interpolation and regression, and they rely on conditioning on input locations. By not providing the temporal index to the decoder, our formulation prevents the global latent from degenerating into a lookup table and instead forces it to encode a genuine, time-agnostic summary of the sequence.

A complementary line of work has focused on learning video representations either through discriminative objectives or more expressive temporal models. Contrastive approaches learn powerful embeddings for downstream tasks such as action recognition and retrieval [125, 33, 93, 37, 117], but do not aim to fully encode the input and therefore do not address compact generative representations. Sequential latent variable models enhanced with NFs similarly aim to reduce redundancy by increasing the flexibility of autoregressive temporal dependencies [108], yet still rely on explicit temporal factorization. In parallel, the extensive literature on video autoencoders and latent video generators has largely pursued spatiotemporal compression through multiple latents, hierarchical structures, or explicit motion modeling, including tokenizer-based generation [166], 3D latent diffusion [58, 174], and recent video VAEs [177, 53, 158, 70]. While some of these approaches incorporate global elements to promote temporal coherence, information remains fundamentally distributed across multiple latent variables and explicit temporal modeling is required. In contrast,

our approach removes temporal modeling from the decoder and shifts the burden of capturing sequence structure entirely to a single global latent, yielding a compact representation that captures the dominant information of the video while leaving only minimal residual variability at the frame level.

5.3 Global Representation Framework

The core assumption of this work is that a single *global* latent element can encode not only the static content of a sequence, but also substantial information about its dynamics. Although motion unfolds over time, global properties of that motion can be summarized into a compact representation that conditions all frames of the sequence.

To make this intuition more explicit, consider a simplified setting in which each frame corresponds to a point in two-dimensional space. A sequence is therefore a trajectory

$$\boldsymbol{x}_{1:T} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T), \qquad \boldsymbol{x}_t = (x_t, y_t) \in \mathbb{R}^2, \tag{5.1}$$

which can be interpreted as a path in the plane. This example is visualized in Fig.5.2.

In conventional sequential modeling, some frames must be encoded *independently*. These play the role of intra-frames: they are modeled without reference to other time steps. If trajectories may explore the plane freely, then, in the absence of additional information, such frames must be described by a distribution that covers all admissible positions. In the extreme case where paths can go anywhere, this leads to a distribution close to uniform over the spatial domain, reflecting maximal uncertainty. Other frames can be encoded *relative* to previously encoded ones. This corresponds to inter-frame modeling. For instance, one may use an autoregressive factorization so that each position is predicted from past positions. This reduces uncertainty locally, since knowledge of $\boldsymbol{x}_{t-1}$ constrains the likely values of $\boldsymbol{x}_t$.

Now suppose that we introduce a global latent variable $\boldsymbol{s}$ inferred from the entire sequence and that conditions every frame. The role of $\boldsymbol{s}$ is to encode information about the full trajectory, that is the global path structure. This improves both types of modeling. For frames that must be encoded

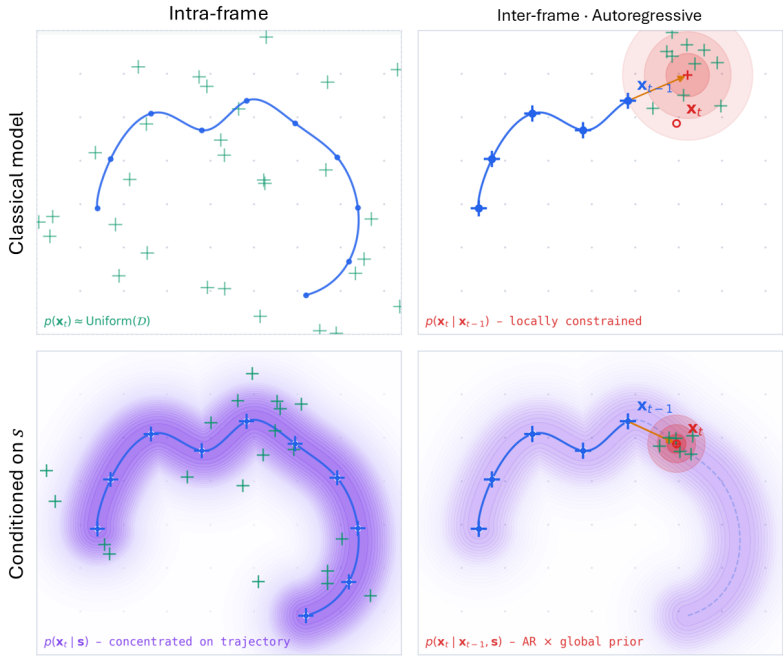


Fig. 5.2 2D temporal path illustration (Sec.5.3). Columns: frame-independent (left) vs. autoregressive (right). Rows: classical formulation (top) vs. conditioning on global $\boldsymbol{s}$ (bottom). Conditioning on $\boldsymbol{s}$ reduces frame-level entropy in both cases.

independently of previous ones, conditioning on $\boldsymbol{s}$ drastically reduces uncertainty. Instead of modeling $\boldsymbol{x}_t$ with a distribution close to uniform over the entire domain, we model it relative to the specific trajectory summarized by $\boldsymbol{s}$. The support of the conditional distribution is therefore much more concentrated. For frames encoded autoregressively, conditioning on $\boldsymbol{s}$ also provides additional information. Even though $\boldsymbol{x}_{t-1}$ constrains $\boldsymbol{x}_t$ locally, it does not necessarily determine the global direction or long-range structure of the path. Since $\boldsymbol{s}$ has "seen" all frames during inference, it carries information about the overall motion and thus refines the conditional distribution of $\boldsymbol{x}_t$ beyond what past frames alone can provide.

Overall, the idea can be interpreted as modeling the sequence with respect to a *learned* 'frame' of reference that has already seen the entire sequence and summarizes both spatial and temporal information into a single element.

Going beyond this simple two-dimensional example, let $\boldsymbol{x}_{1:T}$ be a video sequence with $\boldsymbol{x}_t \in \mathbb{R}^{C\times H\times W}$. We introduce a global static latent variable $\boldsymbol{s} \in \mathcal{S}$ inferred from the full sequence, and we model each frame as

$$\boldsymbol{x}_t = h_\theta(\boldsymbol{\tau}_t; \boldsymbol{s}), \tag{5.2}$$

where $h : \mathcal{S} \times \mathcal{T} \to \mathbb{R}^{C\times H\times W}$ is a measurable function and $\boldsymbol{\tau}_t \in \mathcal{T}$ is a frame-dependent latent variable. Crucially, h operates independently for each frame t: no temporal modeling is required in the decoder.

In this formulation, $\boldsymbol{s}$ is shared across all time steps and is responsible for encoding all information that is common to the sequence. This includes:

- *static information*, such as objects, background, and scene layout,
- *global dynamic information*, such as the type of motion, the action being performed, and the overall temporal structure of the sequence.

The variables $\boldsymbol{\tau}_t$ then capture only the remaining variability required to reconstruct each specific frame given $\boldsymbol{s}$. In other words, $\boldsymbol{\tau}_t$ represents the *residual* once all shared spatio-temporal information has been factored out through $\boldsymbol{s}$.

In the ideal regime where $\boldsymbol{s}$ captures the full content and motion of the sequence, the role of $\boldsymbol{\tau}_t$ becomes minimal and highly structured. It can be interpreted as a learned coordinate describing the position of $\boldsymbol{x}_t$ within the sequence. In that sense, $\boldsymbol{\tau}_t$ behaves almost like a learned notion of "time step" or progression index, but defined in a data-driven manner rather than imposed a priori. The video is effectively encoded in $\boldsymbol{s}$ alone, without requiring any temporal modeling that can be factored out of the representation learning problem entirely. The $\boldsymbol{\tau}_t$ remain available as lightweight structured residuals for downstream tasks that do require sequential modeling, but they are not needed to obtain a compact global descriptor of the sequence.

Therefore, modeling $\boldsymbol{x}_t$ through $h(\boldsymbol{\tau}_t, \boldsymbol{s})$ removes redundancies across frames, both static and dynamic, and isolates only the frame-specific component in $\boldsymbol{\tau}_t$. The global latent variable $\boldsymbol{s}$ acts as a compact summary of the entire sequence, while the temporal evolution is captured through structured residual variations relative to this shared reference. Overall, this allows to recover a desirable compact global representation of video data, directly usable for downstream tasks, while avoiding the need for complex temporal modeling. Such a representation could naturally lead to several potential advantages in:

Generative modeling. Generating $\boldsymbol{s}$ reduces to image-like generation, which is substantially more mature than full video generation [131, 127]. Temporal dynamics can then be modeled as structured minimal variations around this global reference, promoting strong temporal consistency and simplifying modeling.

Compression. Since $\boldsymbol{s}$ is inferred from the full sequence, it acts as a global learned frame of reference that captures shared structure across frames. Encoding $\boldsymbol{s}$ once and conditioning all frames upon it reduces redundancy over time. In terms of density modeling, conditioning on $\boldsymbol{s}$ sharpens the conditional distributions of the frames, potentially enabling more efficient video compression.

Representation learning. Being fully unsupervised, the method can be trained on large-scale unlabeled video corpora. The representation $\boldsymbol{s}$ could then be used as a compact, static descriptor for downstream tasks such as classification, retrieval, or clustering, thereby avoiding the need to model temporal dependencies during supervised fine-tuning.

In the next section, we describe how this conceptual framework can be implemented in practice and how the global latent variable $\boldsymbol{s}$ can be learned effectively from data.

5.4 Method

Using the notations introduced previously where $\boldsymbol{x}_{1:T} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_T)$ denotes a video sequence, the goal is to model each frame as

$$\boldsymbol{x}_t = h_\theta(\boldsymbol{\tau}_t; \boldsymbol{s}), \tag{5.3}$$

where $\boldsymbol{s} \in \mathcal{S}$ is a global static latent variable inferred from the entire sequence, $\boldsymbol{\tau}_t \in \mathcal{T}$ is a frame-dependent latent variable, and h_θ is invertible with respect to $\boldsymbol{\tau}_t$.

The global latent variable $\boldsymbol{s}$ is obtained through an encoder that factorizes spatial and temporal processing. Concretely, stacked layers alternate between shared spatial operators applied independently to each frame, and temporal attention mechanisms operating across time. This interleaved design allows spatial features to be extracted per frame while progressively integrating temporal dependencies. Such a factorization of spatial and temporal interactions yields an efficient architecture, as spatial operators

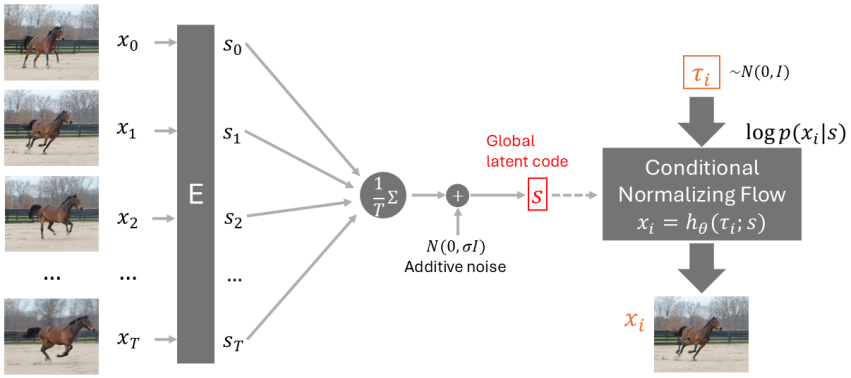


Fig. 5.3 Schematic overview of the proposed model. A sequence encoder extracts per-frame representations $\boldsymbol{s}_t$ which are averaged to produce a global latent $\boldsymbol{s}$. Each frame is modeled conditionally using a NF, producing residual latents $\boldsymbol{\tau}_t$. The flow is invertible, enabling exact likelihood computation and lossless information decomposition into global and frame-wise components.

handle high-dimensional within-frame structure while attention operates only along the temporal dimension [11]. At the final layer, we obtain per-frame latent representations $\boldsymbol{s}_t$, each already enriched with full-sequence context through the temporal attention layers, which are then aggregated by averaging to produce the final global static latent variable $\boldsymbol{s}$. Following Garnelo *et al.* [48], averaging provides a parameter-free, permutation-equivariant aggregation, ensuring a compact fixed-size sequence-level representation while retaining information integrated across time. The complete details on the architecture are provided in App.C.3.

The conditional mapping $h_\theta(\boldsymbol{\tau}_t; \boldsymbol{s})$ is implemented using a scalable transformer based Normalizing Flow [175], modified to condition on the image-shaped latent $\boldsymbol{s}$ via cross-attention. For fixed $\boldsymbol{s}$, the transformation

$$\boldsymbol{\tau}_t = f_\theta(\boldsymbol{x}_t; \boldsymbol{s}) \tag{5.4}$$

is bijective in $\boldsymbol{x}_t$, and h_θ denotes its inverse. The conditional log-likelihood is thus computed exactly through the change-of-variables formula.

$$\log p_\theta(\boldsymbol{x}_t \mid \boldsymbol{s}) = \log p(\boldsymbol{\tau}_t) + \log \left| \det \frac{\partial f_\theta(\boldsymbol{x}_t; \boldsymbol{s})}{\partial \boldsymbol{x}_t} \right|. \tag{5.5}$$

Assuming a standard Gaussian prior for $\boldsymbol{\tau}_t$, the joint conditional likelihood

factorizes as

$$\log p_\theta(\boldsymbol{x}_{1:T} \mid \boldsymbol{s}) = \sum_{t=1}^{T} \log p_\theta(\boldsymbol{x}_t \mid \boldsymbol{s}). \tag{5.6}$$

Because the decoder is invertible and lossless: any information in $\boldsymbol{x}_t$ that is not explained by $\boldsymbol{s}$ must be encoded in $\boldsymbol{\tau}_t$. Hence, $\boldsymbol{\tau}_t$ can be interpreted as the residual information required to reconstruct $\boldsymbol{x}_t$ given the global latent variable.

Since NFs operate in continuous space whereas images are discrete, as standard practice in flow-based density estimation [36, 64], uniform dequantization noise is added to the input frames during training. This transforms the discrete data distribution into a continuous one, allowing the model to learn a valid likelihood by modeling a noise-augmented density. The latent variable $\boldsymbol{s}$ is constrained to lie within the same bounded range as the frames to avoid arbitrarily high-variance latent space, and the same fixed-variance noise is added to $\boldsymbol{s}$ to ensure consistency with the continuous modeling framework. Both the encoder and conditional NF are learned simultaneously. The model is summarized in Fig.5.3.

Theoretical Justification. The reason the proposed model encourages the latent $\boldsymbol{s}$ to encode as much information as possible about the sequence can be understood through the lens of Variational Autoencoder theory [176], by comparing the behavior of the standard VAE formulation with the specific setting induced by our model.

Overall, the model described above can be interpreted as a VAE in which the decoder is a conditional NF. The encoder can be denoted by $q_\phi(\boldsymbol{s} \mid \boldsymbol{x}_{1:T})$, with fixed encoder variance (implemented via additive noise), and the prior $p(\boldsymbol{s})$ is uniform over the bounded support of $\boldsymbol{s}$. The expected evidence lower bound (ELBO) over the data distribution reads:

$$\mathcal{L} = \mathbb{E}_{p_{\text{data}}(\boldsymbol{x}_{1:T})} \left[\mathbb{E}_{q_\phi(\boldsymbol{s}|\boldsymbol{x}_{1:T})} \log p_\theta(\boldsymbol{x}_{1:T} \mid \boldsymbol{s}) - \mathrm{KL}\big(q_\phi(\boldsymbol{s} \mid \boldsymbol{x}_{1:T}) \,\|\, p(\boldsymbol{s})\big) \right] \tag{5.7}$$

The regularization term can be further decomposed as:

$$\mathbb{E}_{p_{\text{data}}(\boldsymbol{x}_{1:T})} \mathrm{KL}\big(q_\phi(\boldsymbol{s} \mid \boldsymbol{x}_{1:T}) \,\|\, p(\boldsymbol{s})\big) = I_q(\boldsymbol{x}_{1:T}; \boldsymbol{s}) + \mathrm{KL}\big(q_\phi(\boldsymbol{s}) \,\|\, p(\boldsymbol{s})\big), \tag{5.8}$$

where $I_q(\boldsymbol{x}_{1:T}; \boldsymbol{s})$ denotes the mutual information under the joint distribution $q_\phi(\boldsymbol{x}_{1:T}, \boldsymbol{s})$, and the second term enforces matching between the aggregated posterior $q_\phi(\boldsymbol{s})$ and the prior $p(\boldsymbol{s})$.

This decomposition makes explicit that, in a *classical* VAE, the ELBO

regularization penalizes the mutual information $I_q(\boldsymbol{x}_{1:T}; \boldsymbol{s})$. In the infinite-capacity regime, the optimization may reduce to minimizing $\mathrm{KL}(q_\phi(\boldsymbol{s}) \| p(\boldsymbol{s}))$, whose optimum is achieved when $I_q(\boldsymbol{x}_{1:T}; \boldsymbol{s}) = 0$. This phenomenon is commonly referred to as the *information preference property* or *posterior collapse* [29, 176]: when the decoder is sufficiently expressive, the model can ignore the latent variable and rely solely on the decoder to model the data distribution, while forcing the posterior to match the prior.

However, in contrast, the proposed formulation does not optimize a KL regularization term that constrains the entropy of the global latent variable. The prior $p(\boldsymbol{s})$ is fixed (uniform over a bounded domain), and the encoder variance is fixed through additive noise. Consequently, there is no learnable mechanism that encourages a reduction of the entropy of $\boldsymbol{s}$ during training. The optimization is therefore driven solely by maximization of the exact conditional log-likelihood:

$$\mathcal{L} = \mathbb{E}_{p_{\text{data}}(\boldsymbol{x}_{1:T})} \mathbb{E}_{q_\phi(\boldsymbol{s}|\boldsymbol{x}_{1:T})} \log p_\theta(\boldsymbol{x}_{1:T} \mid \boldsymbol{s}). \tag{5.9}$$

Assuming sufficient capacity of the conditional NF, the optimal decoder satisfies $p^*_\theta(\boldsymbol{x}_{1:T} \mid \boldsymbol{s}) = q_\phi(\boldsymbol{x}_{1:T} \mid \boldsymbol{s})$, yielding

$$\begin{aligned} \mathbb{E}_{q_\phi(\boldsymbol{x}_{1:T}, \boldsymbol{s})} \log p^*_\theta(\boldsymbol{x}_{1:T} \mid \boldsymbol{s}) &= -H_q(\boldsymbol{x}_{1:T} \mid \boldsymbol{s}) \\ &= I_q(\boldsymbol{x}_{1:T}; \boldsymbol{s}) - H(\boldsymbol{x}_{1:T}). \end{aligned} \tag{5.10}$$

Since $H(\boldsymbol{x}_{1:T})$ is constant with respect to the model parameters,maximizing the likelihood is equivalent to maximizing mutual information $I_q(\boldsymbol{x}_{1:T}; \boldsymbol{s})$. Hence, a maximally informative latent representation is actively promoted.

Unlike classical VAEs, where the ELBO may drive the model toward posterior collapse by penalizing mutual information, the absence of an adaptive KL regularization term here prevents the entropy of $\boldsymbol{s}$ from being minimized during training. The global latent variable $\boldsymbol{s}$ is therefore encouraged to encode as much sequence-level information as possible in order to optimize the reconstruction term $\log p_\theta(\boldsymbol{x}_{1:T} \mid \boldsymbol{s})$.

Finally, because the decoder is invertible, any information in $\boldsymbol{x}_t$ that is not captured by $\boldsymbol{s}$ must necessarily be represented in the residual latent variable $\boldsymbol{\tau}_t$. The model thus naturally decomposes each sequence into a maximally informative global component $\boldsymbol{s}$ and frame-wise residuals $\boldsymbol{\tau}_t$, without suffering from posterior collapse. In this sense, the proposed formulation explicitly promotes an information-rich global representation.

Why a Normalizing Flow Decoder? A crucial design choice in our framework is the use of a conditional NF as decoder. If a standard AE decoder were used, it would take the global latent $\boldsymbol{s}$ and attempt to reconstruct each frame via a pixel-space loss. However, since the same latent $\boldsymbol{s}$ is shared across all frames, this will typically result in a blurry average of the frames and fails to capture temporal variability. In contrast, a conditional NF models the full conditional distribution $\log p_\theta(\boldsymbol{x}_t \mid \boldsymbol{s})$ rather than a pointwise reconstruction. This ensures that $\boldsymbol{s}$ captures global, sequence-level structure, while any information not explained by $\boldsymbol{s}$ is encoded in the residual latent $\boldsymbol{\tau}_t$.

5.5 Experiments

In this section, we empirically validate the central hypothesis of this work, namely that a *single* global latent variable $\boldsymbol{s}$ can encode the essential information of a video sequence.

We structure the evaluation along two complementary axes. First, we analyze the *informational content* of $\boldsymbol{s}$ and verify that it provides a compact yet expressive summary of the sequence, with direct implications for representation learning and generative modeling. Second, we evaluate the proposed formulation from a *density estimation* perspective, highlighting improvements in likelihood and discussing consequences for compression.

We conduct experiments on a diverse set of synthetic and real-world video datasets, including **MHAD** [25], **MUG** [3], **KTH** [135], **BAIR** [42], **Moving MNIST** [142], and **LPCSprites** [128]. Complete details on the implementation, datasets, and evaluation protocols are provided in the Appendix.

5.5.1 Informative global latent $\boldsymbol{s}$

As explained earlier, the proposed model can be viewed as a video VAE with a *single* global latent representation $\boldsymbol{s}$ for the entire sequence.The decoder is a NF and is therefore invertible, yielding exact reconstruction given $(\boldsymbol{s}, \boldsymbol{\tau}_t)$. The natural way to assess the information captured by $\boldsymbol{s}$ alone is therefore to sample from the decoder conditional distribution $p_\theta(\boldsymbol{x}_{1:T} \mid \boldsymbol{s})$.

As shown in Fig.5.4, samples generated from a fixed $\boldsymbol{s}$ resemble frames from the original sequence. They preserve global structure, appearance,

and motion, while differing only in local details. This indicates that $\boldsymbol{s}$ encodes the essential content of the sequence, while $\boldsymbol{\tau}_t$ captures only minimal, frame-specific variations, effectively indexing frames within the sequence. This is precisely the desired behavior: the sequence is compressed into a single maximally informative latent, without temporal modeling, while residual information remains limited. We now provide quantitative results to support this observation.

Evaluation protocol. Most datasets considered provide sequence-level labels, which we exploit to evaluate the discriminative content of the learned representations. We train linear classifiers on top of the global latent $\boldsymbol{s}$ and on the collection of residual latents $\boldsymbol{\tau}_{1:T}$ to predict both *dynamic labels* (e.g., action classes) and *static labels* (e.g., identity or scene attributes when available). Higher classification accuracy indicates that the corresponding representation retains discriminative information about the sequence.

To further quantify class separability, we compute the *silhouette score*, which measures the cohesion and separation of clusters , as well as the *Fisher discriminant ratio*, defined as the ratio of inter-class variance to intra-class variance. For interpretability, we report a normalized version given by $R_{Fisher}/(1 + R_{Fisher})$, which takes values in $[0, 1]$, with higher values indicating better class separation.

We also assess *label consistency under sampling*. Given a sequence, we fix its latent $\boldsymbol{s}$, sample frames from the conditional decoder, and evaluate whether the predicted labels of the generated samples match those of the original sequence. We report the corresponding classification accuracy.

Beyond label-based evaluation, we introduce label-free metrics to assess the intrinsic informativeness of $\boldsymbol{s}$. Inspired by [69, 7], we define the *discriminative score* that directly reflects how informative and discriminative the global latent $\boldsymbol{s}$ is with respect to the sequence. Given a sequence $\boldsymbol{x}_{1:T}$ and its associated latent $\boldsymbol{s}$, we define

$$d_{\text{score}} = \mathbb{E}_{p(\boldsymbol{x}_{1:T})} \frac{p_\theta(\boldsymbol{x}_{1:T} \mid \boldsymbol{s})}{p_\theta(\boldsymbol{x}_{1:T} \mid \boldsymbol{s}) + \sum_{i=1}^{K} p_\theta(\boldsymbol{x}_{1:T} \mid \boldsymbol{s}'_i)}, \tag{5.11}$$

where $\{\boldsymbol{s}'_i\}_{i=1}^{K}$ are latent variables inferred from other sequences. This score takes values in $[0, 1]$ and can be interpreted as the probability that the model selects the correct conditioning latent $\boldsymbol{s}$ over K negative candidates. In other words, it measures how often the model assigns higher likelihood to the true latent compared to alternative ones. This metric is

Table 5.1 Label-based evaluation metrics across datasets.

		MUG		MHAD		KTH		LPCSprites		M-MNIST	
Metric	**Mode**	s	τ	s	τ	s	τ	s	τ	s	τ
Linear Accuracy	dyn	0.7902	0.4863	0.8185	0.5222	0.6593	0.2927	0.9718	0.4316	-	-
	static	1.0000	0.5588	1.0000	0.5729	0.7683	0.4155	1.0000	0.4935	1.0000	0.4291
KNN Accuracy	dyn	0.8510	0.5647	0.7889	0.5111	0.6187	0.2967	0.9271	0.3750	-	-
	static	1.0000	0.5608	1.0000	0.5690	0.8067	0.3548	1.0000	0.4474	1.0000	0.3943
Fisher ratio	dyn	0.7621	0.6463	0.9148	0.6413	0.6413	0.5930	-	-	-	-
	static	0.9257	0.6503	0.9632	0.6461	0.7301	0.6109	-	-	0.9783	0.5854
Silhouette	dyn	0.4258	0.2543	0.4261	0.1795	0.3109	0.0829	-	-	-	-
	static	0.7162	0.2410	0.7988	0.2360	0.6612	0.1876	-	-	0.8264	0.2137
Label Transfer		0.9650		0.8550		0.9500		1.0000		1.0000	

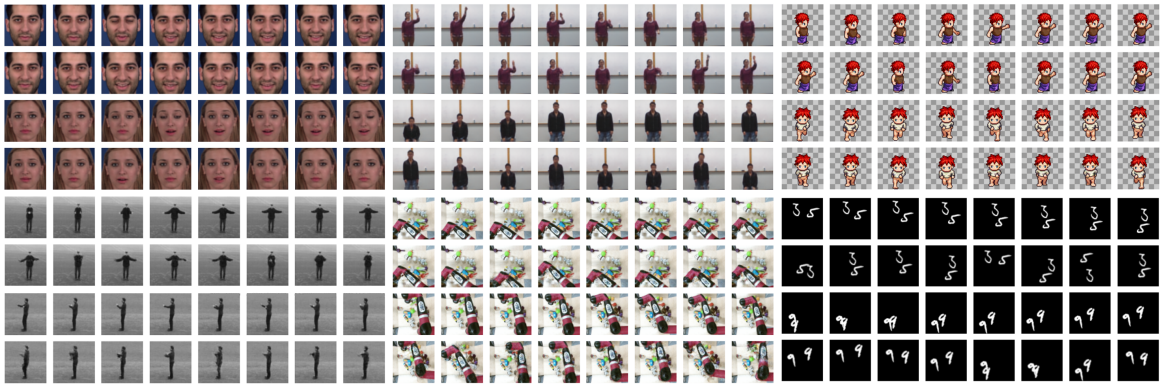


Fig. 5.4 Samples from our model. Odd rows: ground-truth frames. Even rows: random samples from $p_\theta(\boldsymbol{x}_t|\boldsymbol{s})$ given fixed $\boldsymbol{s}$. Samples resemble frames from the original sequence, with no temporal ordering as frames are modeled independently.

Table 5.2 Label-free evaluation metrics across datasets.

Dataset	$d_{score} \uparrow$	dist.-$d_{score} \uparrow$	$R_\tau \downarrow$
MUG	0.9123	0.9682	0.2650
MHAD	0.8342	0.9624	0.2315
KTH	0.7235	0.9083	0.3334
BAIR	0.9857	0.9565	0.2995
LPCSprites	1.0000	0.9880	0.1439
M-MNIST	0.9623	0.9746	0.1952

closely related to contrastive objectives [114], and admits a direct Mutual Information interpretation. In particular, $\log d_{\text{score}}$ provides a lower bound on $I(\boldsymbol{x}_{1:T}; \boldsymbol{s})$ (see App.C.2.2). We compute two variants: one using the exact NF likelihood, and one based on distances between generated samples and the target sequence.

Finally, we quantify the relative contribution of $\boldsymbol{s}$ and $\boldsymbol{\tau}_t$ to the variability of the reconstruction through a sensitivity analysis of the decoder h. Denoting the Jacobians $J_{\boldsymbol{\tau}} = \frac{\partial h}{\partial \boldsymbol{\tau}}, J_{\boldsymbol{s}} = \frac{\partial h}{\partial \boldsymbol{s}}$, we define the ratio

$$R_{\boldsymbol{\tau}} = \frac{\mathbb{E}\left[\|J_{\boldsymbol{\tau}}\|_F^2\right]}{\mathbb{E}\left[\|J_{\boldsymbol{\tau}}\|_F^2\right] + \mathbb{E}\left[\|J_{\boldsymbol{s}}\|_F^2\right]}, \tag{5.12}$$

which measures the fraction of output variability explained by the residual latents. A low value of $R_{\boldsymbol{\tau}}$ indicates that most variability is captured by the global latent $\boldsymbol{s}$.

Results. Overall, the results are consistent with the qualitative observations. For label-based metrics (Tab.5.1), linear classifiers trained on $\boldsymbol{s}$ achieve high accuracy across all datasets, while classifiers on $\boldsymbol{\tau}_{1:T}$ perform significantly worse, showing that $\boldsymbol{s}$ captures both static and dynamic information. Performance slightly decreases on fine-grained or ambiguous actions (e.g., walking vs. jogging in KTH, or directional motions in MHAD), which is expected as such distinctions rely on precise temporal cues that may be partially compressed in a global representation. The Fisher discriminant and silhouette scores further confirm that $\boldsymbol{s}$ induces a well-structured latent space with strong class separability, unlike $\boldsymbol{\tau}_t$. Label transfer results also show high consistency: samples generated from a fixed $\boldsymbol{s}$ retain the original labels and remain close in feature space, indicating that $\boldsymbol{\tau}_t$ captures only limited residual variability.

For label-free metrics (Tab.5.2), the same conclusions hold. The discriminative score consistently attains high values, indicating that the model strongly favors the correct latent $\boldsymbol{s}$ over alternatives. This provides direct evidence of high mutual information between $\boldsymbol{s}$ and $\boldsymbol{x}_{1:T}$, and conversely low informative content in $\boldsymbol{\tau}_t$. Consistently, the Jacobian-based analysis shows that only a small fraction of the output variability is explained by the residuals $\boldsymbol{\tau}_t$.

Overall, the decomposition induced by the model is clearly observed: $\boldsymbol{s}$ encodes the dominant structure of the sequence, while $\boldsymbol{\tau}_t$ accounts for fine-grained, frame-specific details and effectively acts as a low-dimensional coordinate indexing variations conditioned on $\boldsymbol{s}$. These results demonstrate that a single global latent variable can summarize a video sequence while retaining most of its informative content. This makes $\boldsymbol{s}$ a compact and static representation well suited for downstream tasks, without requiring explicit temporal modeling. Furthermore, since $\boldsymbol{s}$ determines global content and dynamics, video generation reduces to sampling a single image-like latent followed by low-complexity sampling of $\boldsymbol{\tau}_t$, naturally promoting temporal consistency.

5.5.2 Density estimation

We now evaluate whether the proposed representation leads to improvements in density estimation, thereby validating the intuition developed in Sec.5.3. In particular, we investigate whether conditioning on the global latent variable $\boldsymbol{s}$ reduces the uncertainty of the data distribution.

Since the model operates on discrete data, dequantization noise is added during training, allowing us to model a continuous density. This enables the computation of likelihoods with respect to the original discrete inputs $\tilde{\boldsymbol{x}}$ via

$$\tilde{p}(\tilde{\boldsymbol{x}}) = \int_{\boldsymbol{\epsilon} \in [0,\text{bin}]^D} p_\theta(\tilde{\boldsymbol{x}} + \boldsymbol{\epsilon})\, d\boldsymbol{\epsilon}, \tag{5.13}$$

from which we report results in bits-per-dimension (BPD), following the standard literature approach [36, 175].

We consider two complementary settings. First, we evaluate frame-wise likelihoods without temporal dependencies, corresponding to an intra-frame encoding regime. Second, we introduce an autoregressive (AR) prior over the residual latents $\boldsymbol{\tau}_t$, allowing the model to exploit temporal dependencies across frames, similarly to [108]. This corresponds to an inter-frame setting. For conditional models, we evaluate density by accounting

Table 5.3 BPD results. *Independent*: average BPD over all frames under $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ prior, with (w/ $\boldsymbol{s}$) and without (w/o $\boldsymbol{s}$) the global latent. *Autoregressive*: average BPD over frames under the learned temporal prior.

	Independent		*Autoregressive*	
Dataset	w/o $\boldsymbol{s}$	w/ $\boldsymbol{s}$	w/o $\boldsymbol{s}$	w/ $\boldsymbol{s}$
BAIR	3.568	2.599	2.735	2.371
KTH	2.324	2.045	1.956	1.827
LPC Sprites	2.132	2.041	1.764	1.602
MHAD	2.147	1.985	1.911	1.798
MUG	2.520	2.344	2.164	2.031

for the joint likelihood $p_\theta(\boldsymbol{s}, \boldsymbol{x}_{1:T})$, thereby including the cost of encoding the global latent $\boldsymbol{s}$ (=ELBO on $\log p(x_{1:T})$, see App.C.2.3).

As our goal is to isolate the effect of introducing a global latent variable, we use comparable architectures across settings and focus on the relative impact of conditioning.

Results. The results are presented in Tab.5.3. As shown previously, the global latent $\boldsymbol{s}$ captures most of the information in the sequence. As a result, conditioning on $\boldsymbol{s}$ consistently improves likelihood across all datasets and configurations. Intuitively, the sequence-level information is encoded once in $\boldsymbol{s}$, and each frame is then modeled relative to this shared reference, leading to significantly improved likelihood.

In the frame-independent setting, conditioning provides systematic gains, as $\boldsymbol{s}$ captures shared information such as scene structure, object identity, and global motion patterns. Conditioning on $\boldsymbol{s}$ therefore reduces the support of the conditional distribution and sharpens the likelihood of each frame. In the autoregressive setting, the unconditional model already reduces uncertainty as past frames are observed, reflecting temporal redundancy in video data. However, the conditional model consistently achieves lower BPD. This is because $\boldsymbol{s}$ provides global information about the sequence, including long-range structure and motion, which complements the local information captured by the autoregressive model. By providing this global latent, most of the information does not need to be modeled by the temporal prior, leaving only residual temporal variations to capture and thereby easing the temporal modeling task.

Overall, these results confirm that the global latent $\boldsymbol{s}$ reduces the burden on both spatial and temporal modeling, leading to improved likelihood

and a simpler decomposition of the data distribution.

Information-theoretic interpretation. The negative log-likelihood (Tab. 5.3) corresponds to the cross-entropy between the data distribution and the model. The improvement obtained by conditioning on $\boldsymbol{s}$ can therefore be interpreted as a reduction in conditional uncertainty. In particular, it reflects how much information about $\boldsymbol{x}_t$ is captured by the global latent. A formal connection with mutual information and compression is provided in App.C.2.3. In the autoregressive setting, the observed gap further indicates that $\boldsymbol{s}$ captures information beyond what is explained by past frames alone. Encoding $\boldsymbol{s}$ once and conditioning all frames upon it reduces temporal redundancy, yielding a more efficient factorization of the data distribution. Hence, while designing a full compression system is beyond the scope of this work, our results indicate that the proposed representation could be a promising direction for future video compression models.

5.6 Conclusion

In this work, we addressed the fundamental question of whether a video sequence can be summarized by a single compact representation. Departing from classical approaches that distribute information across time, we proposed a framework in which a global latent variable s encodes the entire sequence and serves as a shared reference for all frames. This formulation shifts the role of temporal modeling from the decoder to the representation itself, enabling a fixed-size, image-like encoding of video data. Our results show that such a representation is not only feasible but highly effective: s captures the dominant structure of the sequence while inducing well-separated and informative latent spaces, and conditioning on s consistently improves density estimation across all datasets and configurations.

Beyond these empirical findings, this work suggests a different perspective on video representation learning: rather than increasing model complexity to handle temporal dependencies, one can instead seek representations that absorb this structure into a compact latent variable. This opens several promising directions for future research, including scaling to large and diverse video corpora for general-purpose unsupervised representations, as well as applications in generative modeling and video compression, where encoding shared sequence-level information once could yield substantially more efficient models.

6

Conclusion and Perspectives

This thesis began with a diagnostic. After the landmark impossibility result of Locatello *et al.* [103], the field of disentangled representation learning found itself in an uncomfortable position: a decade of practical progress was shown to rest on foundations that could not, in principle, be theoretically justified. The standard response was to introduce more supervision (labels, pairs, or explicit causal graphs), or to design ever more elaborate inductive biases while accepting the lack of guarantees. The working hypothesis of this thesis was different: that the impossibility is not a wall but a signpost, pointing toward a definition of "factor" that is more faithful to how factors actually work in data. If the standard definition is the source of the impossibility, then the right response is not to add supervision until the definition becomes tractable, but to ask whether the definition itself is the right one. That question turned out to be generative. It opened a line of work in which theory is not a post-hoc justification for empirical choices but the primary driver of them: each modeling decision in this thesis follows from a formal argument, and each formal argument was motivated by something that needed explaining in the data. What connects the three contributions that result is the role that theory plays throughout, not as a retrospective guarantee for decisions already made, but as the engine that drives them.

What follows records what was found, what was achieved, and what remains to be done.

6.1 What This Thesis Has Established

Three contributions make up this thesis, each self-contained and each adapted from a peer-reviewed publication. What this section attempts is not a summary of their technical content, which the preceding chapters have covered in full, but a reckoning with what they collectively establish: why the results matter beyond the specific problems they address, and what the field looks like differently in light of them.

6.1.1 A new definition of what a factor is

The central claim running through this work is that the field of disentangled representation learning has, for some time, been working from an incomplete notion of what a factor is. The classical formulation defines factors through their statistical relationships: either as statistically independent components, following the tradition of Independent Component Analysis, or, in more recent causal approaches, as nodes in an explicit causal graph whose dependencies are modeled and constrained. Both families of approaches locate the meaning of a factor in how it relates to other factors, whether through distributional orthogonality or through causal structure. This thesis argues that neither is sufficient. The functional perspective advanced relocates the definition entirely. A factor is meaningful not because of its statistical relationship to the others, nor because it occupies a node in a causal graph, but because it acts on the observations through a distinct and separable generative mechanism, independently of how the factors happen to be jointly distributed in any particular dataset. This shift is not merely semantic: it is the move that unlocks identifiability in the unsupervised setting without causal commitments, and it resolves in a principled way the failures that the independence-based impossibility result had left unexplained.

6.1.2 Sequential disentanglement: functional separation without independence

In the sequential setting of Chap.3, the functional perspective has immediate and consequential empirical implications. Static and dynamic factors are generically causally dependent: the static content of a scene shapes

the possible range of its dynamics, yet the two act through separate mechanisms that the proposed formalism makes precise. Enforcing statistical independence between the two codes, as all prior sequential disentanglement methods do, directly contradicts this causal structure and produces entangled representations. This is not a subtle failure: it is a structural contradiction between the modeling assumption and the data-generating process, and it manifests as systematic degradation on every correlated benchmark examined. The asymmetric architecture and shuffle constraint proposed in this thesis enforce functional separation instead, and Prop.3.3 establishes that all disentanglement conditions are jointly satisfied by minimizing the resulting Evidence Lower Bound, without any auxiliary loss or dataset-specific component. The outcome is the first provably disentangled sequential model that handles dependent factors, achieved without any supervision.

6.1.3 Identifiability from functional orthogonality

Ch.4 generalizes this insight to the static-data setting, formalizing functional independence as an orthogonality condition on the columns of the generative Jacobian. Under this condition, the generative process becomes identifiable up to the natural ambiguities of permutation and element-wise reparameterisation, even when the latent factors are statistically dependent, provided the data covers all combinations of factor values. This result does not contradict the impossibility of Locatello *et al.* [103] but resolves it: the impossibility holds precisely when factorhood is defined statistically, and it dissolves when factorhood is defined functionally.

The theoretical framework also furnishes, as a by-product, the first rigorous mechanistic explanation for the empirical success of VAEs at disentanglement. The factorized prior and the diagonal approximate posterior jointly enforce the dual constraint (global independence pressure and local Jacobian orthogonality) that the theory identifies as necessary and sufficient for identifiability. Normalizing Flows, sharing the same independence pressure but lacking the orthogonality bias, consistently fail to disentangle, as predicted theoretically and confirmed experimentally. This explanation is not merely retrospective: it clarifies which architectural choices are load-bearing and which are incidental, and it points directly toward architectures that could inherit the disentanglement properties of VAEs without their known limitations.

6.1.4 From disentanglement to compact global video representations

Ch.5 explores what happens when the disentanglement constraint is deliberately relaxed. Removing the shuffle operation from the sequential architecture leaves the global latent free to absorb arbitrary sequence-level information; under maximum-likelihood training with a fixed-variance encoder, this drives the global latent to be maximally informative about the entire sequence, avoiding posterior collapse by design. The result is a compact, fixed-size video descriptor that captures both static content and the global structure of the dynamics without any explicit temporal modeling in the decoder. Beyond its direct utility for representation learning, this chapter demonstrates the broader applicability of the architectural principles developed for disentanglement: the same structural choices that enforce factorization when constrained produce rich, holistic representations when the constraint is lifted.

6.1.5 A unifying thread: theory and practice in dialogue

A thread running through all three contributions is the productive relationship between what can be proved and what can be computed. Prop.3.3, 4.1, and 4.2 are not exercises in mathematical generality for its own sake: each one either directly motivates a modeling choice or provides a principled explanation for a phenomenon already visible in the data. The proof that minimizing the shuffled Evidence Lower Bound is equivalent to satisfying the disentanglement conditions leads directly to the model architecture. The proof that Jacobian orthogonality yields identifiability explains the gap in disentanglement performance between Normalizing Flows and VAEs. The proof that the unconstrained objective maximizes mutual information explains why the global latent does not collapse. In each case, the mathematics does not merely justify what was already done; it reveals what should be done next.

6.2 Limitations and Perspectives

Every theoretical contribution is only as strong as the assumptions it rests on, and every proof of concept only as convincing as the scale at which it has been tested. The discussions that follow take this seriously, distin-

guishing what has been established from what remains to be done, and identifying the directions where further investigation is most likely to be fruitful.

6.2.1 Functional Orthogonality: Open Questions at the Frontier

Chap.4 is the theoretical heart of this thesis, and it is where the deepest open questions reside. The framework is rigorous within its stated assumptions, and the empirical results on controlled synthetic benchmarks closely track the theoretical predictions. The honest assessment, however, is that these benchmarks are precisely that: controlled. Scaling the framework to realistic, high-dimensional settings raises challenges that are both practical and foundational.

The most fundamental open question is whether the data-generating processes that matter in practice (natural images, real video, biological signals) admit decompositions satisfying Asm.4.1. The orthogonal Jacobian condition is proposed as a definitional criterion for what a meaningful factor is, grounded in the Independent Causal Mechanisms principle [134] and supported by prior works [52, 129]. As discussed at length in App.B.1.4, both assumptions are empirically and theoretically motivated rather than arbitrary. Yet the question of whether nature, or the portion of it captured in a finite dataset, actually satisfies these conditions closely enough for the framework to yield meaningful representations at scale cannot be answered by theory alone. It requires large-scale empirical investigation beyond the scope of this thesis, and it connects to a deeper epistemological question: is functional orthogonality a property of the world, a property of human cognition, or a modeling convenience that happens to align with the structure that learning algorithms find tractable? The framework does not resolve this question; it provides the right language to ask it precisely.

A related practical challenge concerns verification. In the synthetic experiments of Chap.4, the assumptions hold by construction and ground-truth latent factors are available for evaluation. In realistic settings, neither is true: one cannot directly inspect the Jacobian of nature's generative process, and there are no ground-truth factors to recover. Developing diagnostic tools that can assess, from observations alone, whether a learned representation approximately satisfies the functional orthogonality conditions is an important open problem and a prerequisite for deploying the

framework beyond controlled benchmarks.

Perhaps the most promising concrete direction opened by Chap.4 is what it describes as the path beyond VAEs. The theoretical framework is model-agnostic: identifiability follows from structural properties of the generative mapping, not from any particular architecture. The orthogonal Jacobian bias could therefore be transplanted into Normalizing Flows, Diffusion models, or large-scale autoregressive architectures, which avoid the posterior collapse, blurry reconstructions, and information-preference tradeoffs that limit VAEs. The current bottleneck is computational: enforcing Jacobian orthogonality exactly is expensive, and the existing regularization methods [52] do not yet scale to the high-dimensional settings that matter. Developing scalable approximations, whether through structured architectures whose Jacobians are orthogonal by construction, stochastic estimators of the orthogonality penalty, or architecture-specific efficient implementations, represents one of the clearest and most consequential directions opened by this work, and one where the theoretical guarantees established here provide a well-defined target.

6.2.2 Identifiability as a Compass, Not a Destination

A further perspective concerns the status of identifiability itself, and it is perhaps the one most easily misread. The thesis is built around identifiability results, and it would be natural to conclude that strict identifiability is the objective toward which everything is directed. It is not. Identifiability is better understood as a compass than as a destination: its value lies less in the guarantee it furnishes for any particular trained model than in what it reveals about *why* disentanglement succeeds or fails, and in the design principles it thereby licenses. Real data is messy: the data-generating processes encountered in practice will rarely satisfy all assumptions exactly and ground-truth factors are unavailable. Insisting on strict identifiability as a precondition for useful representation learning would therefore be self-defeating; what one actually wants is a representation that separates the meaningful directions of variation well enough to be useful for downstream reasoning, transfer, generation, or compression, whether or not it coincides with the ground truth to the letter.

Seen this way, the contribution of the identifiability analysis is *diagnostic and prescriptive* rather than certifying. It tells us which structural properties of a model are load-bearing for disentanglement. None of these ex-

planations requires that any deployed model be exactly identifiable; they require only that the theory correctly locates the mechanism. The orthogonality criterion is a clear instance: even when the generative process does not satisfy Asm.4.1 closely enough to guarantee recovery of the true factors, encouraging an orthogonal generative Jacobian still biases the model toward separating the directions of variation that act on the observations through distinct modes. The representation that results may recover no canonical or human-nameable concept, yet it will tend to decouple aspects of the data that are functionally independent, and for many purposes this functional decoupling is exactly what is wanted. Identifiability points toward the bias; the bias is useful in its own right.

This reframing also opens a concrete direction for future work: the study of what happens when the identifiability conditions are not met exactly, but only *partially*, as real data will typically only approximately satisfy them. The thesis did not have the scope to pursue this question, but one might speculate that partial satisfaction of the conditions yields a correspondingly partial identifiability: rather than recovering each factor up to permutation and element-wise reparameterization, one would recover the latent space only up to a decomposition into *subspaces*, disentangled across blocks but still entangled within them. The map between the learned and true latent variables, which full identifiability forces to be a permutation composed with a coordinate-wise reparameterization, might then become block-wise instead. Such a regime would interpolate between full disentanglement, when every block is one-dimensional, and complete non-identifiability, when a single block spans the entire latent space, and it would arguably be a more faithful model of real data, where some groups of factors are cleanly separable while others remain irreducibly intertwined. Characterizing it is a natural continuation of the functional program initiated here.

6.2.3 Global Video Representations: A Promising Direction Still in Its Infancy

Chap.5 occupies a different position from the preceding contributions. It is deliberately exploratory: a proof of concept for a new way of thinking about video representation rather than a mature system. The central empirical finding is robust (a single global latent can effectively summarize a video sequence, and conditioning on it consistently improves density es-

timation across all tested configurations), and the theoretical justification for why the model avoids posterior collapse is clean. But the experiments are conducted on datasets of moderate scale and complexity, and several dimensions of the problem remain essentially unexplored.

The most immediate open question is one of scale. The six datasets used for evaluation are well-controlled and relatively short, and it is not yet known whether the concentration of video information in a single latent variable persists as sequence length grows, as scene complexity increases, or as the diversity of content expands toward the kind of open-domain video found in large video corpora. These are empirical questions that can be addressed given sufficient compute, and they constitute the natural next experimental program.

An equally important gap is the absence of established baselines and benchmarks for the specific task proposed in this chapter. The problem of summarizing an entire video into a single compact latent without explicit temporal modeling in the decoder is new enough that no standard points of comparison yet exist. Establishing evaluation protocols and releasing the model for community comparison would allow the strengths and limitations of the approach to be assessed more precisely, and would help organize a nascent research direction.

Of the downstream applications motivated by Chap.5, compression is perhaps the most tantalizing. The density estimation results already show that conditioning all frames on a shared $\boldsymbol{s}$ reduces bits-per-dimension across all tested settings. This reflects a principled factorization of the information in a video sequence: sequence-level redundancy is absorbed into $\boldsymbol{s}$ and encoded once, while each frame is decoded relative to this common reference with only the residual variation that $\boldsymbol{s}$ does not explain. Connecting this factorization to the engineering of modern learned video codecs and evaluating it at codec-level compression metrics would complete the picture. The generative direction is equally compelling: if $\boldsymbol{s}$ is an image-like object encoding both static content and global dynamics, video generation reduces to sampling $\boldsymbol{s}$ from a prior (a problem in image-like generation that the field has mastered at impressive scale) and decoding each frame as a lightweight residual variation. Given more time and computational resources, this is the direction where continued investment seems most likely to yield substantial returns, and it is where this research would have gone next.

6.3 Final Thoughts

These years of research have been, above all, an exercise in learning to ask better questions. The field of disentangled representation learning entered a period of productive uncertainty after the impossibility result, and this thesis is one attempt to find a way forward: not by circumventing the impossibility through supervision, but by reconsidering the definition that makes it unavoidable. The functional perspective developed here, grounding factorhood in the geometry of generative influence rather than in statistical relationships, has led to results that are, I believe, both theoretically principled and empirically meaningful. The standard account of disentanglement fails on three fronts simultaneously: it lacks theoretical guarantees, it fails on correlated data, and it cannot explain why the methods that work in practice actually work. The framework proposed here addresses all three, at least within the settings examined. It also opened, somewhat unexpectedly, a third direction: that the same architectural choices carry well beyond disentanglement, toward compact global video representations with real potential for generation and compression.

Whether these contributions will prove durable as the field scales is a question I am genuinely excited to see answered. The path toward larger models, richer data, and more expressive generative architectures is well underway, and I hope the theoretical foundations laid here, the functional definition of factors, the identifiability results, the global video representation, will find their place in that larger story. These years of research have opened more questions than they have closed, which I have come to regard as the clearest sign that the work was pointed in a worthwhile direction.

Despite the inevitable constraints of time and compute that shape every doctoral project, I believe this work has brought modest but hopefully useful contributions to the understanding of what disentanglement is, why it sometimes works, and how it can be made to work on purpose. I look forward to seeing where the questions it leaves open will lead.

A

Supplementary Material for Chapter 3

A.1 Proofs and Detailed Derivations

A.1.1 Derivation of the Per-Sequence and Dataset ELBO

The Evidence lower bound (ELBO) provides a lower estimate of the marginal data log-likelihood $\log p(\boldsymbol{x}_{1:T})$. Given the joint distribution $p(\boldsymbol{x}_{1:T}, \boldsymbol{f})$ and the approximate posterior distribution $q(\boldsymbol{f}|\boldsymbol{x}_{1:T})$, the ELBO can be computed as:

$$\log p(\boldsymbol{x}_{1:T}) \tag{A.1}$$

$$\geq \mathbb{E}_{q(\boldsymbol{f}|\boldsymbol{x}_{1:T})} \log \frac{p(\boldsymbol{x}_{1:T}, \boldsymbol{f})}{q(\boldsymbol{f}|\boldsymbol{x}_{1:T})} \tag{A.2}$$

$$= \mathbb{E}_{q(\boldsymbol{f}|\boldsymbol{x}_{1:T})} \log \frac{p(\boldsymbol{f})p(\boldsymbol{x}_{1:T}|\boldsymbol{f})}{q(\boldsymbol{f}|\boldsymbol{x}_{1:T})} \tag{A.3}$$

$$= \mathbb{E}_{q(\boldsymbol{f}|\boldsymbol{x}_{1:T})} \log p(\boldsymbol{x}_{1:T}|\boldsymbol{f}) - KL(q(\boldsymbol{f}|\boldsymbol{x}_{1:T})||p(\boldsymbol{f})) \tag{A.4}$$

$$= \mathbb{E}_{q(\boldsymbol{f}|\boldsymbol{x}_{1:T})} \log p(\boldsymbol{\lambda}_{1:T}|\boldsymbol{f}) + \log \left| \det \frac{\delta \boldsymbol{h}^{-1}(\boldsymbol{x}_{1:T}; \boldsymbol{f})}{\delta \boldsymbol{x}_{1:T}} \right| - KL(q(\boldsymbol{f}|\boldsymbol{x}_{1:T})||p(\boldsymbol{f})) \tag{A.5}$$

where in the last line we model $p(\boldsymbol{x}_{1:T}|\boldsymbol{f})$ using a NF, $\boldsymbol{x}_{1:T} = \boldsymbol{h}(\boldsymbol{\lambda}_{1:T}; \boldsymbol{f})$. This proof is adapted from the standard VAE framework [29, 7]. This expression constitutes the per-sequence ELBO.

However, in Ch.3, we consider two important modifications from that original ELBO. We replace $p(\boldsymbol{\lambda}_{1:T}|\boldsymbol{f})$ with $p(\boldsymbol{\lambda}_{1:T})$ and condition the NF on the shuffled static vectors instead of the aggregated one. We show how these modifications still lead to a valid lower bound. We start with the dataset ELBO and further note that, because the frame static vectors are extracted independently, $q(\boldsymbol{f}_{1:T}|\boldsymbol{x}_{1:T}) = \prod_t q(\boldsymbol{f}_t|\boldsymbol{x}_t) = q(\boldsymbol{f}|\boldsymbol{x}_{1:T})$:

$$\mathbb{E}_{p_{Data}} \log p(\boldsymbol{x}_{1:T}) \tag{A.6}$$

$$\geq \mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} \log p(\boldsymbol{\lambda}_{1:T}|\boldsymbol{f}_{1:T}) + \log \left| \det \frac{\delta \boldsymbol{h}^{-1}(\boldsymbol{x}_{1:T}; \boldsymbol{f}_{1:T})}{\delta \boldsymbol{x}_{1:T}} \right|$$

$$- KL(q(\boldsymbol{f}|\boldsymbol{x}_{1:T})||p(\boldsymbol{f})) \tag{A.7}$$

$$= \mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} \log p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T}) - KL(q(\boldsymbol{f}|\boldsymbol{x}_{1:T})||p(\boldsymbol{f})) \tag{A.8}$$

Let us denote $\boldsymbol{f}_{\pi_T}$, the randomly shuffled $\boldsymbol{f}_{1:T} \sim q(\boldsymbol{f}_{1:T}|\boldsymbol{x}_{1:T})$. The ELBO can extended as:

$$\text{Eq.A.8} = \mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} \log p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{\pi_T}) + \underbrace{\log \frac{p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T})}{p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{\pi_T})}}_{(*1)}$$

$$- KL(q(\boldsymbol{f}|\boldsymbol{x}_{1:T})||p(\boldsymbol{f})) \tag{A.9}$$

We can note from Eq.A.8 that, to maximize the ELBO (with q fixed), the optimum value of the likelihood $\mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} \log p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T})$ is achieved for $KL(q(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T})||p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T}))$ minimized i.e., $p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T}) = q(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T})$. As a result we have:

$$\mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} \log p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T}) = \mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} \log q(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T})$$

$$- KL(\log q(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T})|| \log p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T})) \tag{A.10}$$

$$\Leftrightarrow \mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} \log p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T}) \geq \mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} \log q(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T}) \tag{A.11}$$

$$\Leftrightarrow \mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} \log p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{\pi_T}) + \log \frac{p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T})}{p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{\pi_T})}$$

$$\geq \mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} \log q(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T}) \tag{A.12}$$

$$\Leftrightarrow \mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} \log \frac{p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T})}{p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{\pi_T})} \geq \mathbb{E}_{q(\boldsymbol{f}_{1:T})} KL(q(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T})||p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{\pi_T})) \tag{A.13}$$

with equality when $p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T}) = q(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T})$ and thus

$$\boxed{(*1) = \mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} \log \frac{p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T})}{p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{\pi_T})} \geq \mathbb{E}_{q(\boldsymbol{f}_{1:T})} KL(q(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T})||q(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{\pi_T})) \geq 0} \tag{A.14}$$

Injecting this into the ELBO gives:

$$\begin{aligned}
\text{Eq.A.9} &= \mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} \log p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{\pi_T}) + \log \frac{p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T})}{p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{\pi_T})} \\
&\quad - KL(q(\boldsymbol{f}|\boldsymbol{x}_{1:T})||p(\boldsymbol{f})) && \text{(A.15)} \\
&\geq \mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} \log p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{\pi_T}) - KL(q(\boldsymbol{f}|\boldsymbol{x}_{1:T})||p(\boldsymbol{f})) && \text{(A.16)} \\
&= \mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} \log p(\boldsymbol{\lambda}_{1:T}|\boldsymbol{f}_{\pi_T}) + \log \left| \det \frac{\delta \boldsymbol{h}^{-1}(\boldsymbol{x}_{1:T}; \boldsymbol{f}_{\pi_T})}{\delta \boldsymbol{x}_{1:T}} \right| \\
&\quad - KL(q(\boldsymbol{f}|\boldsymbol{x}_{1:T})||p(\boldsymbol{f})) && \text{(A.17)} \\
&= \mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} \log p(\boldsymbol{\lambda}_{1:T}) + \underbrace{\log \frac{p(\boldsymbol{\lambda}_{1:T}|\boldsymbol{f}_{\pi_T})}{p(\boldsymbol{\lambda}_{1:T})}}_{(*2)} \\
&\quad + \log \left| \det \frac{\delta \boldsymbol{h}^{-1}(\boldsymbol{x}_{1:T}; \boldsymbol{f}_{\pi_T})}{\delta \boldsymbol{x}_{1:T}} \right| - KL(q(\boldsymbol{f}|\boldsymbol{x}_{1:T})||p(\boldsymbol{f})) && \text{(A.18)}
\end{aligned}$$

As done previously, Eq.A.16 is maximized (with q fixed) for $KL(q(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T})||$ $p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{\pi_T}))$ minimized i.e., $q(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T}) = p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{\pi_T})$. We thus have:

$$\begin{aligned}
&\mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} \log p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{\pi_T}) = \mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} \log q(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T}) \\
&\quad - KL(q(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T})||p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{\pi_T})) && \text{(A.19)} \\
\Leftrightarrow &\mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} \log p(\boldsymbol{\lambda}_{1:T}) + \log \frac{p(\boldsymbol{\lambda}_{1:T}|\boldsymbol{f}_{\pi_T})}{p(\boldsymbol{\lambda}_{1:T})} + \log \left| \det \frac{\delta \boldsymbol{h}^{-1}}{\delta \boldsymbol{x}_{1:T}} \right| \\
&\quad \geq \mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} \log q(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T}) && \text{(A.20)}
\end{aligned}$$

$$\Leftrightarrow \mathbb{E}_{q(\boldsymbol{f}_{1:T},\boldsymbol{x}_{1:T})} \log \frac{p(\boldsymbol{\lambda}_{1:T}|\boldsymbol{f}_{\pi_T})}{p(\boldsymbol{\lambda}_{1:T})} \geq \mathbb{E}_{q(\boldsymbol{f}_{1:T},\boldsymbol{x}_{1:T})} - \log \left| \det \frac{\delta \boldsymbol{h}^{-1}}{\delta \boldsymbol{x}_{1:T}} \right| + KL(q(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T})||p(\boldsymbol{\lambda}_{1:T})) \tag{A.21}$$

with equality when $q(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T}) = p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{\pi_T})$ meaning

$$\mathbb{E}_{q(\boldsymbol{f}_{1:T},\boldsymbol{x}_{1:T})} \log \frac{p(\boldsymbol{\lambda}_{1:T}|\boldsymbol{f}_{\pi_T})}{p(\boldsymbol{\lambda}_{1:T})} \geq \mathbb{E}_{q(\boldsymbol{f}_{1:T}),p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{\pi_T})} - \log \left| \det \frac{\delta \boldsymbol{h}^{-1}}{\delta \boldsymbol{x}_{1:T}} \right| + KL(p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{\pi_T})||p(\boldsymbol{\lambda}_{1:T})) \tag{A.22}$$

$$\Rightarrow \boxed{(*2) = \mathbb{E}_{q(\boldsymbol{f}_{1:T},\boldsymbol{x}_{1:T})} \log \frac{p(\boldsymbol{\lambda}_{1:T}|\boldsymbol{f}_{\pi_T})}{p(\boldsymbol{\lambda}_{1:T})} \geq \mathbb{E}_{q(\boldsymbol{f}_{1:T})} KL(p(\boldsymbol{\lambda}_{1:T}|\boldsymbol{f}_{\pi_T})||p(\boldsymbol{\lambda}_{1:T})) \geq 0} \tag{A.23}$$

As a result the ELBO becomes:

$$\begin{aligned} \text{Eq.A.18} &= \mathbb{E}_{q(\boldsymbol{f}_{1:T},\boldsymbol{x}_{1:T})} \log p(\boldsymbol{\lambda}_{1:T}) + \log \frac{p(\boldsymbol{\lambda}_{1:T}|\boldsymbol{f}_{\pi_T})}{p(\boldsymbol{\lambda}_{1:T})} \\ &\quad + \log \left| \det \frac{\delta \boldsymbol{h}^{-1}(\boldsymbol{x}_{1:T}; \boldsymbol{f}_{\pi_T})}{\delta \boldsymbol{x}_{1:T}} \right| - KL(q(\boldsymbol{f}|\boldsymbol{x}_{1:T})||p(\boldsymbol{f})) \qquad \text{(A.24)} \\ &\geq \mathbb{E}_{q(\boldsymbol{f}_{1:T},\boldsymbol{x}_{1:T})} \log p(\boldsymbol{\lambda}_{1:T}) + \log \left| \det \frac{\delta \boldsymbol{h}^{-1}(\boldsymbol{x}_{1:T}; \boldsymbol{f}_{\pi_T})}{\delta \boldsymbol{x}_{1:T}} \right| \\ &\quad - KL(q(\boldsymbol{f}|\boldsymbol{x}_{1:T})||p(\boldsymbol{f})) \qquad \text{(A.25)} \end{aligned}$$

which is the loss optimized by our proposed model. **Overall, we demonstrated that the two proposed modifications still lead to a valid lower bound on the marginal data log-likelihood**. In App.A.1.4, we show how that novel modified ELBO does naturally encourage disentanglement.

A.1.2 Proof of Proposition 3.1

Proposition 3.1. *Consider the generative process and conditions presented in Sec.3.3. Further assume that $dim(\boldsymbol{f}) = n_f \geq n_s$. Let $\boldsymbol{x}$ and $\boldsymbol{x}^*$ denote two frames (or disjoint sets of frames) belonging to the same sequence $\boldsymbol{x}_{1:T}$ and $\mathbb{D}$ be a divergence between two distributions. Given unlimited data from $p(\boldsymbol{x}_{1:T})$, any candidate posterior distribution $q(\boldsymbol{f}|\boldsymbol{x}_{1:T})$ that minimizes*

$$L_{prop1} = \mathbb{E}_{p(\boldsymbol{x}_{1:T})} \mathbb{D}(q(\boldsymbol{f}|\boldsymbol{x}), q(\boldsymbol{f}|\boldsymbol{x}^*)) - I(\boldsymbol{f}, \boldsymbol{s}) + I(\boldsymbol{s}, \boldsymbol{s}), \tag{A.26}$$

is disentangled in the sense that $q(\boldsymbol{f}) = \int q(\boldsymbol{f}|\boldsymbol{x}_{1:T})p(\boldsymbol{x}_{1:T})d\boldsymbol{x}_{1:T}$*, the aggregated posterior, is a reparametrization of* $p(\boldsymbol{s})$*.*

Proof. Based on the generative process, the true data conditional likelihood $p(\boldsymbol{x}_t|\boldsymbol{z}_t)$ is fully specified by the smooth and invertible function $\boldsymbol{g}$. Consequently, the true posterior $p(\boldsymbol{s}|\boldsymbol{x}_t)$ is determined by $\boldsymbol{g}^{-1}_{1:n_s}$, i.e., the inverse of $\boldsymbol{g}$ restricted to the first n_s dimensions, for any t. Hence, It is assumed that the model family for the candidate posterior distribution $q(\boldsymbol{f}|\boldsymbol{x})$ is specified by the set of smooth functions $\boldsymbol{g}' : \mathcal{X} \to \mathcal{F}$ with $\mathcal{F} \subseteq \mathbb{R}^{n_f}$ and $\boldsymbol{x}$ a frame or set of frames from a sequence $\boldsymbol{x}_{1:T}$, $\boldsymbol{f} = \boldsymbol{g}'(\boldsymbol{x})$. To prove identifiability, we have to show that, for the $\boldsymbol{g}'$ associated to $q(\boldsymbol{f}|\boldsymbol{x})$, $\boldsymbol{g}'(\boldsymbol{x}) = \boldsymbol{m}(\boldsymbol{s})$ with $\boldsymbol{m} : \mathcal{S} \to \mathcal{F}$ a smooth and invertible function with smooth inverse.

One could note that Prop.3.1 is similar to what has been proposed to provably extract the content for content-style disentanglement in Thm.4.4 of [151]. This is not surprising since, as uncovered in the main text, static-dynamic disentanglement is actually a content-style problem generalized to multiple temporally varying styles. As a result, both problems are equivalent and the codes are defined similarly. In fact, since the global code $\boldsymbol{s}$ is in both cases extracted as the common part, Thm.4.4 in [151] is also directly valid to extract the static factors for static-dynamic disentanglement. However, compared to the aforementioned theorem, Prop.3.1 extends Thm.4.4 of [151] to stochastic representations and, by leveraging mutual information, does not require the number of static variables n_s to be known apriori or impose a uniform distribution to the variables. Nonetheless, as it directly extends the work of Von Kügelgen *et al.* [151], the proof can be structured similarly.

The proof is constructed as follows : First we show that any candidate function that attains the global minimum of $\mathcal{L}_{\text{prop1}}$ outputs a representation that has maximum mutual information with $\boldsymbol{s}$ and is invariant across inputs of the same sequence i.e., $\boldsymbol{x}$ and $\boldsymbol{x}^*$. Second, leveraging the similarities of static-dynamic disentanglement with content-style disentanglement [14, 151], we use results from [151] to show that the obtained representation does not depend on the dynamic variables $\boldsymbol{d}_{1:T}$ and relate to $\boldsymbol{s}$ through a smooth function. Finally, we show that the latter function is invertible using the mutual information properties.

Step 1. As $\mathbb{D}$ is a divergence its minimum value is $\mathbb{D}(q(\boldsymbol{f}|\boldsymbol{x}), q(\boldsymbol{f}|\boldsymbol{x}^*)) = 0$ when both distributions are equal i.e., $\boldsymbol{g}'(\boldsymbol{x}) = \boldsymbol{g}'(\boldsymbol{x}^*)$. Moreover, from

the properties of mutual information $I(\boldsymbol{f}, \boldsymbol{s}) \leq I(\boldsymbol{s}, \boldsymbol{s})$ with equality when $\boldsymbol{f}$ capture all information in $\boldsymbol{s}$ ($n_f \geq n_s$). The global minimum of $\mathcal{L}_{\text{prop1}} = 0$ is achieved when the mutual information is maximized.

Consider the function $\boldsymbol{b} = \boldsymbol{v} \circ \boldsymbol{g}_{1:n_s}^{-1} : \mathcal{X} \to \mathcal{F}$, the composition of $\boldsymbol{g}_{1:n_s}^{-1}$ described above and $\boldsymbol{v} : \mathcal{S} \to \mathcal{F}$ a smooth and invertible function with smooth inverse. $\boldsymbol{b}$, being a composition of two smooth invertible functions, is smooth and invertible. We can show that the function $\boldsymbol{b}$ attains the global minimum of $\mathcal{L}_{\text{prop1}}$:

$$\boldsymbol{b}(\boldsymbol{x}) = \boldsymbol{v}(\boldsymbol{s}) = \boldsymbol{b}(\boldsymbol{x}^*) \Rightarrow \mathbb{D}(q(\boldsymbol{f}|\boldsymbol{x}), q(\boldsymbol{f}|\boldsymbol{x}^*)) = 0, \tag{A.27}$$

$$I(\boldsymbol{b}(\boldsymbol{x}), \boldsymbol{s}) = I(\boldsymbol{v}(\boldsymbol{s}), \boldsymbol{s}) = I(\boldsymbol{s}, \boldsymbol{s}) \tag{A.28}$$

with Eq.A.27 resulting from the fact that $\boldsymbol{x}$ and $\boldsymbol{x}^*$ are frames from the same sequence and Eq.A.28 resulting from the invariance of MI to invertible transformations. This ensures the existence of a smooth function that attains the global minimum.

Let us now consider **any** smooth function $\boldsymbol{g}' : \mathcal{X} \to \mathcal{F}$ that achieves the global minimum of $\mathcal{L}_{\text{prop1}}$:

$$\mathcal{L}_{\text{prop1}} = \mathbb{E}_{p(\boldsymbol{x}_{1:T})} \mathbb{D}(q(\boldsymbol{f}|\boldsymbol{x}), q(\boldsymbol{f}|\boldsymbol{x}^*)) - I(\boldsymbol{f}, \boldsymbol{s}) + I(\boldsymbol{s}, \boldsymbol{s}) = 0 \tag{A.29}$$

All candidate functions can be written as $\boldsymbol{g}' = \boldsymbol{m} \circ \boldsymbol{g}^{-1}$ with $\boldsymbol{m} : \mathcal{Z} \to \mathcal{F}$ a smooth function that maps the true frame code $\boldsymbol{z}$ to $p(\boldsymbol{f})$. Because $\boldsymbol{g}(\boldsymbol{z}) = \boldsymbol{x}$ with $\boldsymbol{z} = (\boldsymbol{s}, \boldsymbol{d})$, Eq.A.29 results in:

$$\mathcal{L}_{\text{prop1}} = 0 \iff \begin{cases} \boldsymbol{m}(\boldsymbol{s}, \boldsymbol{d}) = \boldsymbol{m}(\boldsymbol{s}, \boldsymbol{d}^*) \\ I(\boldsymbol{m}(\boldsymbol{s}, \boldsymbol{d}), \boldsymbol{s}) = I(\boldsymbol{s}, \boldsymbol{s}). \end{cases} \tag{A.30}$$

This means that the representation $\boldsymbol{f} = \boldsymbol{m}(\boldsymbol{s}, \boldsymbol{d})$ must be invariant between frames belonging to the same sequence and have maximum mutual information with $\boldsymbol{s}$.

Step 2. We now show that $\boldsymbol{f} = \boldsymbol{g}'(\boldsymbol{x}) = \boldsymbol{m}(\boldsymbol{s}, \boldsymbol{d})$ does not depend on the true dynamic factors $\boldsymbol{d}$. As explained above, the static and dynamic codes are defined similarly to content/style in [151] and in both cases the same invariance condition is required i.e., $\boldsymbol{m}(\boldsymbol{s}, \boldsymbol{d}) = \boldsymbol{m}(\boldsymbol{s}, \boldsymbol{d}^*)$. We thus directly refer to Step 2 of Thm 4.2 in [151] that proved that, under the invariance condition and the generative process defined in Sec.3.3, $\boldsymbol{f}$ can only depend on the true static factors $\boldsymbol{s}$.

Step 3. Following previous steps, we have that $\boldsymbol{f} = \boldsymbol{m}(\boldsymbol{s})$ with $\boldsymbol{m}$ smooth. It remains to show that the function $\boldsymbol{m}$ is also invertible. From Eq.A.30, the representation must follow:

$$I(\boldsymbol{f}, \boldsymbol{s}) = I(\boldsymbol{m}(\boldsymbol{s}), \boldsymbol{s}) = I(\boldsymbol{s}, \boldsymbol{s}). \tag{A.31}$$

Mutual information is invariant under non-linear homeomorphisms [92] and $I(\boldsymbol{m}(\boldsymbol{s}), \boldsymbol{s}) = I(\boldsymbol{s}, \boldsymbol{s})$ iff $\boldsymbol{m}(\boldsymbol{s})$ is invertible (otherwise $I(\boldsymbol{m}(\boldsymbol{s}), \boldsymbol{s}) < I(\boldsymbol{s}, \boldsymbol{s})$). Hence, we finally obtain $\boldsymbol{f} = \boldsymbol{g}'(\boldsymbol{x}) = \boldsymbol{m}(\boldsymbol{s})$ with $\boldsymbol{m} : \mathcal{S} \to \mathcal{F}$ a smooth and invertible function. By the change of variables formula, we have:

$$q(\boldsymbol{f}) = p(\boldsymbol{m}^{-1}(\boldsymbol{f})) \left| \det \frac{\delta \boldsymbol{m}^{-1}(\boldsymbol{f})}{\delta \boldsymbol{f}} \right|, \tag{A.32}$$

showing that $q(\boldsymbol{f})$ is a reparametrization of $p(\boldsymbol{s})$, concluding the proof. The obtained representation $\boldsymbol{f}$ captures all and only the static factors $\boldsymbol{s}$ and is thus disentangled.

A.1.3 Proof of Proposition 3.2

Proposition 3.2. *Consider the same framework as in Prop.3.1 and the static code $\boldsymbol{f}$ to be disentangled as described above. Further assume that $dim(\boldsymbol{\lambda}) = n_\lambda \geq n_d$. Any smooth and invertible candidate function $\boldsymbol{h}$ s.t. $\boldsymbol{x}_{1:T} = \boldsymbol{h}(\boldsymbol{f}, \boldsymbol{\lambda}_{1:T})$ which minimizes*

$$L_{prop2} = I(\boldsymbol{f}, \boldsymbol{\lambda}_{1:T}), \tag{A.33}$$

is disentangled in the sense that $q(\boldsymbol{\lambda}_{1:T})$ is a reparametrization of $p(\boldsymbol{\epsilon}_{d,1:T})$.

Proof. In order to prove that $\boldsymbol{\lambda}_{1:T}$ identifies the true dynamic exogeneous variables $\boldsymbol{\epsilon}_{d,1:T}$, we start by noting that, following the assumptions and generative process, the frame codes $\boldsymbol{x}_{1:T}$ can be written as

$$\boldsymbol{x}_t = \boldsymbol{g}(\boldsymbol{s}, \boldsymbol{h}'_d(\boldsymbol{s}, \boldsymbol{\epsilon}_{d,<t}, \boldsymbol{\epsilon}_{d,t})) = \boldsymbol{w}(\boldsymbol{s}, \boldsymbol{\epsilon}_{d,\leq t}), \tag{A.34}$$

where $\boldsymbol{w}$ denote an invertible and smooth function with smooth inverse that represents the composition of the causal process and the successive invertible mappings. Hence, by denoting $\boldsymbol{w}' = \boldsymbol{w}^{-1} \circ \boldsymbol{h}$ and considering the

fact that $\boldsymbol{h}$ generates the frames auto-regressively, we obtain from Prop.3.2:

$$\boldsymbol{x}_t = \boldsymbol{h}(\boldsymbol{f}, \boldsymbol{\lambda}_{\leq t}) \tag{A.35}$$

$$\Leftrightarrow \quad (\boldsymbol{s}, \boldsymbol{\epsilon}_{d,\leq t}) = \boldsymbol{w}'(\boldsymbol{f}, \boldsymbol{\lambda}_{\leq t}) \tag{A.36}$$

with $\boldsymbol{w}'$ a smooth and invertible function. There is an invertible mapping between the codes and the ground-truth factors. By assumption, $\boldsymbol{f}$ is disentangled and thus, following the development in Sec.A.1.2, $\boldsymbol{f} = \boldsymbol{m}(\boldsymbol{s})$ with $\boldsymbol{m}$ smooth and invertible. Hence, as claimed in the main text, by using an invertible mapping $\boldsymbol{h}$ to model the frames, if $\boldsymbol{f}$ is disentangled, the dynamic codes $\boldsymbol{\lambda}_{1:T}$ have to be informative and encompass at least $\boldsymbol{\epsilon}_{d,1:T}$ (since $\boldsymbol{w}'$ is invertible). It remains to show that the dynamic codes $\boldsymbol{\lambda}_{1:T}$ are not dependent on the static factors $\boldsymbol{s}$.

By the definition of exogeneous variables we have :

$$I(\boldsymbol{s}, \boldsymbol{\epsilon}_{d,1:T}) = KL(p(\boldsymbol{s}, \boldsymbol{\epsilon}_{d,1:T}) || p(\boldsymbol{s})p(\boldsymbol{\epsilon}_{d,1:T})) = 0. \tag{A.37}$$

Since $\boldsymbol{w}'$ and $\boldsymbol{m}$ are invertible, we can make the change of variables :

$$p(\boldsymbol{s}) = q(\boldsymbol{f}) \left| \det \frac{\delta \boldsymbol{m}(\boldsymbol{s})}{\delta \boldsymbol{s}} \right|, \tag{A.38}$$

$$p(\boldsymbol{s}, \boldsymbol{\epsilon}_{d,1:T}) = q(\boldsymbol{f}, \boldsymbol{\lambda}_{1:T}) \left| \det \frac{\delta \boldsymbol{w}'^{-1}(\boldsymbol{s}, \boldsymbol{\epsilon}_{d,1:T})}{\delta \boldsymbol{s}, \boldsymbol{\epsilon}_{d,1:T}} \right|. \tag{A.39}$$

Eq.A.39 can be further simplified. Indeed, because $\boldsymbol{f}$ is assumed disentangled, it only depends on the static factors $\boldsymbol{s}$ and thus $\boldsymbol{w}'^{-1}(\boldsymbol{s}, \boldsymbol{\epsilon}_{d,1:T})$ restricted to the first n_f output dimensions must be equal to $\boldsymbol{m}$:

$$\boldsymbol{w}'^{-1}(\boldsymbol{s}, \boldsymbol{\epsilon}_{d,1:T})_{1:n_f} = \boldsymbol{m}(\boldsymbol{s}) = \boldsymbol{f}. \tag{A.40}$$

This means the Jacobian term in Eq.A.39 can be factorized as :

$$\left| \det \frac{\delta \boldsymbol{w}'^{-1}(\boldsymbol{s}, \boldsymbol{\epsilon}_{d,1:T})}{\delta \boldsymbol{s}, \boldsymbol{\epsilon}_{d,1:T}} \right| = \left| \det \frac{\delta \boldsymbol{m}(\boldsymbol{s})}{\delta \boldsymbol{s}} \right| \left| \det \frac{\delta \boldsymbol{a}(\boldsymbol{s}, \boldsymbol{\epsilon}_{d,1:T})}{\delta \boldsymbol{\epsilon}_{d,1:T}} \right| \tag{A.41}$$

with $\boldsymbol{a}$ denoting $\boldsymbol{w}'^{-1}$ restricted the last output dimensions corresponding to $\boldsymbol{\lambda}_{1:T}$.

Next, we use the fact that Prop.3.2 imposes the mutual information $I(\boldsymbol{f}, \boldsymbol{\lambda}_{1:T})$ to be minimized. The global minimum is achieved when $I(\boldsymbol{f}, \boldsymbol{\lambda}_{1:T})$ equal 0, meaning $q(\boldsymbol{f}, \boldsymbol{\lambda}_{1:T}) = q(\boldsymbol{f})q(\boldsymbol{\lambda}_{1:T})$. Overall, Eq.A.39 can thus be rewritten as :

$$p(\boldsymbol{s}, \boldsymbol{\epsilon}_{d,1:T}) = q(\boldsymbol{f})q(\boldsymbol{\lambda}_{1:T}) \left|\det \frac{\delta \boldsymbol{m}(\boldsymbol{s})}{\delta \boldsymbol{s}}\right| \left|\det \frac{\delta \boldsymbol{a}(\boldsymbol{s}, \boldsymbol{\epsilon}_{d,1:T})}{\delta \boldsymbol{\epsilon}_{d,1:T}}\right| \tag{A.42}$$

Finally, by injecting Eq.A.38 and A.42 in Eq.A.37 and then simplifying the common terms, we obtain

$$KL(p(\boldsymbol{s}, \boldsymbol{\epsilon}_{d,1:T})||p(\boldsymbol{s})p(\boldsymbol{\epsilon}_{d,1:T})) = 0 \tag{A.43}$$

$$\Leftrightarrow \quad KL(q(\boldsymbol{\lambda}_{1:T}) \left|\det \frac{\delta \boldsymbol{a}(\boldsymbol{s}, \boldsymbol{\epsilon}_{d,1:T})}{\delta \boldsymbol{\epsilon}_{d,1:T}}\right| ||p(\boldsymbol{\epsilon}_{d,1:T})) = 0 \tag{A.44}$$

$$\Leftrightarrow \quad q(\boldsymbol{\lambda}_{1:T}) \left|\det \frac{\delta \boldsymbol{a}(\boldsymbol{s}, \boldsymbol{\epsilon}_{d,1:T})}{\delta \boldsymbol{\epsilon}_{d,1:T}}\right| = p(\boldsymbol{\epsilon}_{d,1:T}) \tag{A.45}$$

with the last equation being the change of variable formula. $\boldsymbol{\lambda}_{1:T}$ does not encompass the static factors $\boldsymbol{s}$ and we obtain that $\boldsymbol{\lambda}_{1:T} = \boldsymbol{a}^{-1}(\boldsymbol{\epsilon}_{d,1:T})$:

$$q(\boldsymbol{\lambda}_{1:T}) = p(\boldsymbol{\epsilon}_{d,1:T}) \left|\det \frac{\delta \boldsymbol{a}^{-1}(\boldsymbol{\lambda}_{1:T})}{\delta \boldsymbol{\lambda}_{1:T}}\right|, \tag{A.46}$$

showing that $q(\boldsymbol{\lambda}_{1:T})$ is a reparametrization of $p(\boldsymbol{\epsilon}_{d,1:T})$, concluding the proof. If the static code $\boldsymbol{f}$ is disentangled, by imposing the independence between the codes, we enforce the Jacobian of $\boldsymbol{w}'$ to be block-diagonal with blocks matching the codes and the true factors.

A.1.4 Proof of Proposition 3.3

Proposition 3.3. *Consider the model proposed in Sec.3.4.2 of the main text and further assume $\alpha \gg \beta$. Minimizing*

$$\mathcal{L} = \mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} - \alpha \left(\log p(\boldsymbol{\lambda}_{1:T}) + \log \left| det \frac{\delta \boldsymbol{h}^{-1}(\boldsymbol{x}_{1:T}; \boldsymbol{f}_{\pi_T})}{\delta \boldsymbol{x}_{1:T}} \right| \right) + \beta KL(q(\boldsymbol{f}|\boldsymbol{x}_{1:T})||p(\boldsymbol{f})) \tag{A.47}$$

is equivalent to minimizing both L_{prop1} and L_{propr2}. Prop.3.1 and 3.2 are satisfied and the learned representation $(\boldsymbol{f}, \boldsymbol{\lambda}_{1:T})$ is disentangled in the sense of Def.3.1.

Proof. We start by noting that $\mathcal{L}$ is the negative dataset ELBO obtained in Eq.A.25 where, following the approach in [60] as it is common for VAE approaches, the likelihood and divergence terms of the VAE have been weighted using different coefficients. We propose to further develop $\mathcal{L}$ as :

$$\begin{aligned}
\mathcal{L} &= \mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} - \alpha \left(\log p(\boldsymbol{\lambda}_{1:T}) + \log \left| \det \frac{\delta \boldsymbol{h}^{-1}(\boldsymbol{x}_{1:T}; \boldsymbol{f}_{\pi_T})}{\delta \boldsymbol{x}_{1:T}} \right| \right) \\
&\quad + \beta KL(q(\boldsymbol{f}|\boldsymbol{x}_{1:T})||p(\boldsymbol{f})) \qquad \text{(A.48)} \\
&= \mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} - \alpha \left(\log p(\boldsymbol{\lambda}_{1:T}|\boldsymbol{f}_{\pi_T}) + \log \frac{p(\boldsymbol{\lambda}_{1:T})}{p(\boldsymbol{\lambda}_{1:T}|\boldsymbol{f}_{\pi_T})} \right. \\
&\quad \left. + \log \left| \det \frac{\delta \boldsymbol{h}^{-1}(\boldsymbol{x}_{1:T}; \boldsymbol{f}_{\pi_T})}{\delta \boldsymbol{x}_{1:T}} \right| \right) + \beta KL(q(\boldsymbol{f}|\boldsymbol{x}_{1:T})||p(\boldsymbol{f})) \qquad \text{(A.49)} \\
&\geq \mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} - \alpha \log p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{\pi_T}) + \alpha KL(p(\lambda_{1:T}|\boldsymbol{f}_{\pi_T})||p(\lambda_{1:T})) \\
&\quad + \beta KL(q(\boldsymbol{f}|\boldsymbol{x}_{1:T})||p(\boldsymbol{f})) \qquad \text{(A.50)} \\
&= \mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} - \alpha \log p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T}) + \alpha \log \frac{p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T})}{p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{\pi_T})} \\
&\quad + \alpha KL(p(\lambda_{1:T}|\boldsymbol{f}_{\pi_T})||p(\lambda_{1:T})) + \beta KL(q(\boldsymbol{f}|\boldsymbol{x}_{1:T})||p(\boldsymbol{f})) \qquad \text{(A.51)} \\
&\geq \mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} - \alpha \log p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T}) + \alpha KL(q(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T})||q(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{\pi_T})) \\
&\quad + \alpha KL(p(\lambda_{1:T}|\boldsymbol{f}_{\pi_T})||p(\lambda_{1:T})) + \beta KL(q(\boldsymbol{f}|\boldsymbol{x}_{1:T})||p(\boldsymbol{f})) \qquad \text{(A.52)}
\end{aligned}$$

The two inequalities result from Eq.A.23 and A.14 subsequently. It can be observed that, compared to the classical VAE loss, the two modifications made i.e, replace $p(\lambda_{1:T}|\boldsymbol{f}_{\pi_T})$ with $p(\lambda_{1:T})$ and condition on the shuffled static vectors $\boldsymbol{f}_{\pi_T}$, cause the apparition of two extra divergence terms $KL(p(\lambda_{1:T}|\boldsymbol{f}_{\pi_T})||p(\lambda_{1:T}))$ and $KL(q(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T})||q(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{\pi_T}))$ respectively. We will now develop the different terms in Eq.A.52. We note that, because $q(\boldsymbol{f}_{1:T}|\boldsymbol{x}_{1:T}) = \prod_t q(\boldsymbol{f}_t|\boldsymbol{x}_t)$ and $\boldsymbol{f}_{\pi_T}$ are the shuffled $\boldsymbol{f}_{1:T}$, we have $q(\boldsymbol{f}_{\pi_T}|\boldsymbol{x}_{1:T}) = q(\boldsymbol{f}_{1:T}|\boldsymbol{x}_{\pi_T})$ and $q(\boldsymbol{f}_{\pi_T}) = q(\boldsymbol{f}_{1:T})$.

1.
$$\begin{aligned}
&\mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} \log p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T}) \\
&= \mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} \log p(\boldsymbol{x}_{1:T}) + \log \frac{p(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T})}{p(\boldsymbol{x}_{1:T})} \\
&\geq \mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} \log p(\boldsymbol{x}_{1:T}) + KL(q(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T})||q(\boldsymbol{x}_{1:T})) \\
&= \mathbb{E}_{q(\boldsymbol{f}_{1:T}, \boldsymbol{x}_{1:T})} \log p(\boldsymbol{x}_{1:T}) + I(\boldsymbol{f}, \boldsymbol{x}_{1:T}) \qquad \text{(A.53)}
\end{aligned}$$

2.
$$\begin{aligned}&\mathbb{E}_{q(\boldsymbol{f}_{1:T})}KL(q(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{1:T})||q(\boldsymbol{x}_{1:T}|\boldsymbol{f}_{\pi_T}))\\&=\mathbb{E}_{q(\boldsymbol{x}_{1:T})}KL(q(\boldsymbol{f}_{1:T}|\boldsymbol{x}_{1:T})||q(\boldsymbol{f}_{\pi_T}|\boldsymbol{x}_{1:T}))\\&=\mathbb{E}_{q(\boldsymbol{x}_{1:T})}KL(q(\boldsymbol{f}_{1:T}|\boldsymbol{x}_{1:T})||q(\boldsymbol{f}_{1:T}|\boldsymbol{x}_{\pi_T}))\end{aligned} \tag{A.54}$$

3.
$$\mathbb{E}_{q(\boldsymbol{f}_{1:T})}KL(p(\lambda_{1:T}|\boldsymbol{f}_{\pi_T})||p(\lambda_{1:T}))\approx I(\boldsymbol{\lambda}_{1:T},\boldsymbol{f}_{\pi_T}) \tag{A.55}$$

4.
$$\mathbb{E}_{q(\boldsymbol{x}_{1:T})}KL(q(\boldsymbol{f}|\boldsymbol{x}_{1:T})||p(\boldsymbol{f}))=KL(q(\boldsymbol{f})||p(\boldsymbol{f}))+I(\boldsymbol{f},\boldsymbol{x}_{1:T}) \tag{A.56}$$

By substituting those expressions in Eq.A.52, we get:

$$\begin{aligned}\text{Eq.A.52}\gtrsim\ &\mathbb{E}_{q(\boldsymbol{f}_{1:T},\boldsymbol{x}_{1:T})}-\alpha\log p(\boldsymbol{x}_{1:T})-(\alpha-\beta)I(\boldsymbol{f},\boldsymbol{x}_{1:T})\\&+\alpha KL(q(\boldsymbol{f}_{1:T}|\boldsymbol{x}_{1:T})||q(\boldsymbol{f}_{1:T}|\boldsymbol{x}_{\pi_T}))\\&+\alpha I(\boldsymbol{\lambda}_{1:T},\boldsymbol{f}_{\pi_T})+\beta KL(q(\boldsymbol{f})||p(\boldsymbol{f}))\end{aligned} \tag{A.57}$$

If we additionally consider that between the frames $\boldsymbol{x}_{1:T}$ only the dynamic variables $\boldsymbol{\epsilon}_{1:T}$ vary, we have for $\mathbb{E}_{q(\boldsymbol{x}_{1:T})}KL(q(\boldsymbol{f}_{1:T}|\boldsymbol{x}_{1:T})||q(\boldsymbol{f}_{1:T}|\boldsymbol{x}_{\pi_T}))\approx I(\boldsymbol{f},\boldsymbol{\epsilon}_{1:T}|\boldsymbol{s})=I(\boldsymbol{f},\boldsymbol{\epsilon}_{1:T})$. Henceforth, we finally have:

$$\begin{aligned}\mathcal{L}\gtrsim\ &\mathbb{E}_{q(\boldsymbol{f}_{1:T},\boldsymbol{x}_{1:T})}-\alpha\log p(\boldsymbol{x}_{1:T})+\beta KL(q(\boldsymbol{f})||p(\boldsymbol{f}))+\underbrace{\alpha I(\boldsymbol{\lambda}_{1:T},\boldsymbol{f}_{\pi_T})}_{L_{prop2}}\\&\underbrace{-(\alpha-\beta)I(\boldsymbol{f},\boldsymbol{s})+\beta\sum_t KL(q(\boldsymbol{f}_t|\boldsymbol{x}_t)||q(\boldsymbol{f}_t|\boldsymbol{x}_{\pi(t)}))}_{L_{prop1}}\end{aligned} \tag{A.58}$$

which incorporates L_{prop1} and L_{prop2} if $\alpha>\beta$. By imposing, during training, $p(\boldsymbol{\lambda}_{1:T}|\boldsymbol{f}_{1:T})=p(\boldsymbol{\lambda}_{1:T})$, the model enforces $\lambda_{1:T}$ to be non informative and thus L_{prop2} while conditioning the cNF using the shuffled static vectors results in L_{prop1}.

Overall, by minimizing $\mathcal{L}$, in addition to learn a representation of the data and the prior as in a classical VAE, we also directly minimize an upper bound on L_{prop1} and L_{prop2}. Because the model parametrizes both the posterior $q(\boldsymbol{f}|\boldsymbol{x}_{1:T})$ and the invertible transformation h, the model satisfies Prop.3.1 and 3.2 and is provably disentangled with the factors being identified. This is achieved through a simple shuffle operation on the static vectors and by imposing $p(\lambda_{1:T})$ after the cNF. This concludes our proof.

A.2 Experimental Setup

A.2.1 Datasets

In this section, we describe the different datasets. A summary of the factors of variations for each dataset is provided in Tab.A.1 and examples of sequences are provided in Fig.A.1.

LPCSprites [128] is a dataset made up of image sequences of animated cartoon characters. Each video is composed of 8 RGB frames at a resolution of 64×64. The dynamic factors include three possible motions: walking, casting spells and slashing. The static factors consist of each character's type and color of body, shirt, pants and hair. This differs from the original dataset as the character identity is not entirely defined by the colors. In addition, compared to previous sequential disentanglement works [99, 7, 112, 10, 180, 56], because it is constant throughout the sequences, the orientation of the characters is also considered to be a static factor. The orientation conditions the motion due to the fact that the action does not appear the same way depending on the side the character is facing (e.g. slashing is made with different hands). We use in total 20000 sequences for training and 5000 for testing.

MUG [3] is a facial expression dataset that comprises video sequences of 52 subjects performing six different expressions: anger, disgust, fear, happiness, sadness and surprise. Each sequence is made up of 50 to 160 frames. Following [7], we randomly sample lenght-15 clips from the original videos and crop the face region using Haar Cascades face detection before resizing to 64×64. There are in total 3429 samples, 25% of which are kept for the test set.

MHAD [25] UTD Multimodal Human Action Dataset is a real world action recognition dataset that comprises 861 video sequences, captured using a Microsoft Kinect camera at resolution of 640×480 pixels, of 8 subjects (4 females and 4 males) performing 27 different actions including waving, sitting, throwing, walking etc. Each subject repeated each action 4 times. Similarly to MUG, we randomly sample length-15 clips from the original videos and crop at the subject position before resizing to 96×96.

MPI3D [51] is a real-world dataset made up of images of a robot arm carrying different 3D printed objects at different positions. The dataset consists of 1036800 RGB images at resolution 64×64, corresponding to all the possible combinations of the following factors: camera position,

Table A.1 Summary of the factors of variations for the datasets. In blue are highligted the static factors that condition the dynamic ones.

Dataset	Static factors	Dynamic factors
LPCSprites	orientation, type+color of body, shirt, pants and hair	action (walking, slashing or casting)
MUG	subject identity	facial expression
MHAD	subject identity	person's pose
MPI3D	camera position, background light, object size/shape/color	vertical and horizontal positions
s-dSprites	object color, shape and size	X-Y position
c-dSprites	object color, shape and size	X-Y position and rotation

background light, horizontal/vertical position, object color/shape/size. Length-16 clips are created by sampling subsequent frames from the dataset. The static factors are the camera position, the background light and the object color/shape/size. The dynamic factors (i.e., the motion) is the horizontal/vertical position that varies with time. The motion is made conditional to some of the static factors. If the arm is carrying a large object, it is allowed to have a smaller maximum speed. If the background light is purple, sea green or salmon, the robot arm is only allowed to move vertically, horizontally or both. Finally, the camera position also influences how the motion appears on the frames. We use in total 20000 sequences for training and 5000 for testing.

dSprites [109] is a synthetic dataset of 2D shapes that comprises 737280 64×64 images, corresponding to all the possible combinations of the following factors: shape, size, orientation, position X, position Y. We also add the color as a supplementary factor of variation. Video clips of length 16 are created by sampling subsequent frames. The static factors encompass the object color, shape and size. For the motion, we use two different variants. For the first one, *s-dSprites*, the orientation is fixed and the X/Y-motion of the objects is sampled independently from the static factors. For the second one, *c-dSprites*, the objects are also allowed to rotate. The rotation is conditioned on the shape due to the different order of rotational symmetry between the different shapes. Moreover, larger objects have lower maximum speeds and bounce on the edges according to their size. Each of the two variants consists of 20000 sequences for training and 5000 for testing.

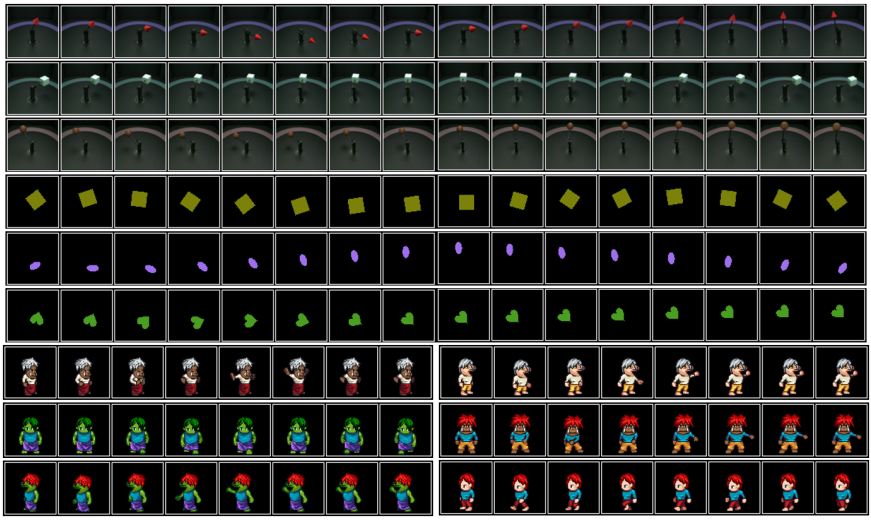


Fig. A.1 Examples of test sequences from *MPI3D* (row 1-3), *c-dSprites* (row 4-6) and *LPCSprites* (row 7-9 in two columns).

A.2.2 Disentanglement Metrics

Accuracy (Acc) measures how well the factors associated with a fixed latent code are preserved when sampling the other codes. A classifier, pretrained with full-supervision on the same train and test sets as the model, predicts the factors based on the reconstructed frames after sampling. For each categorical factor of the dataset, it outputs the probability measures and the predicted labels which are compared to the labels that should be obtained for a perfectly disentangled model.

Inception Score (IS) measures the quality of images created by the generative model. Based on the conditional predicted label distribution $p(y|\boldsymbol{x}_{1:T})$ from the classifier and the marginal predicted label distribution $p(y)$, the inception score is defined as : $IS = \exp(\mathbb{E}_{p(\boldsymbol{x}_{1:T})}[KL(p(y|\boldsymbol{x}_{1:T})||p(y))])$.

Inter-Entropy ($H(y)$) measures the entropy of the marginal predicted label of the generated sequences. Hence, it estimates the diversity for the sampled factors in the generated sequences. It is computed using $p(y)$ obtained from the classifier outputs.

Intra-Entropy ($H(y|x)$) measures the entropy of the conditional predicted label of the generated sequences. Hence, it estimates the confidence of the classifier on the predictions. It is computed using $p(y|\boldsymbol{x}_{1:T})$ obtained

from the classifier outputs.

In the case where the dynamic variables are causally dependent on the static ones, the accuracies for the dynamic factors cannot be computed directly. Indeed, due to the causal relationship, the dynamic factors might change when sampling a new static code. Following the causal model, we can only ensure the dynamic factors are preserved if the conditional static factors are kept fixed. As an example, for the *MPI3D* dataset, if the background is green or purple, the robot arm is only able to move horizontally or vertically respectively. Therefore, if the original sequence has a purple background (move vertically) and the new sampled static factors impose a green background, then the new generated frames should have a different motion and move according to the causal mechanism i.e., horizontally. The model should capture the causal mechanism. Following this remark, to compute the disentanglement metrics for the dynamic factors we thus restrict the samples of the static codes to the ones with the same conditional static factors (i.e., the blue ones in Tab.A.1) and check that the dynamic factors are preserved when sampling the independent static factors.

A.2.3 Architecture and Hyperparameters

The proposed model is implemented in PyTorch [118] and is trained using the Adam optimizer [87] with a learning rate of 0.001. We use a batch size of 64 for all datasets except *MHAD* and *MUG* where we use batches of size 16. The model is trained until convergence for up to 500 epochs.

As it is common practice for VAEs [60], we add coefficients to the different loss terms in order to weight their influence. As an example, by weighting the disentanglement constraints or the KL-divergence terms, we can improve disentanglement and control the information bottleneck of the VAE. This principle is what leads to our provably disentangled model as shown in Sec.A.1.4. In previous static/dynamic sequential disentanglement works [99, 7, 112, 10, 180, 56], these hyperparameters are chosen through grid search and the models that provide the best validation accuracies are used for testing. However, by doing so, the loss coefficients are chosen using the full supervision and the labels of all factors. This is in contradiction with the fact that the method should be self-supervised and labels should only be used for testing. This ultimately results in a bias for the final performances, especially since disentanglement models have been found to be particularly sensitive to the choice of hyperparameters [103].

This problematic goes beyond the subject of this work and is left for future work. Nonetheless, to alleviate this issue, instead of using dataset specific loss coefficients, we propose here to choose coefficients that work well on all datasets. For α and β, we use the following coefficients respectively : 10.0, 0.5.

For the model architecture, the network is composed of the following blocks : Image encoder-decoder, Static encoder and conditional Normalizing Flow.

- **Image encoder-decoder** is in charge of encoding independently each frame into a low-dimensional space $\boldsymbol{x}_{1:T}$ and then reconstructing the videos. The encoder is a convolutional neural network with 5 layers of channels $[64, 128, 256, 512, 128]$. Each layer consists of a convolution with kernel of size 4 followed by a GroupNorm (8 groups) and SiLU activation function. A feature vector of size 128 is outputted for each frame. The decoder is also a convolutional neural network with 5 layers of channels $[512, 256, 128, 64, 3]$. Each layer encompasses a transposed convolution with kernel of size 4 followed by a GroupNorm (8 groups) and SiLu activation function except for the last layer with a Sigmoid. Following [100], this autoencoder is pretrained to encode/decode the frames without disentanglement and Gaussian noise is added on the frame feature vectors $\boldsymbol{x}_{1:T}$ to prevent collapse.

- **Static encoder** takes the frame codes $\boldsymbol{x}_{1:T}$ as input and outputs the static code $\boldsymbol{f}$. It is made of a MLP with 2 linear layers each followed by a SiLU activation function. Two final linear layers predict the mean and standard deviation of the static code. The MLP processes each frame independently and the predictions $\boldsymbol{f}_{1:T}$ are then aggregated following the approach in [14]. All layers are of dimension 128 and $n_f = 128$. The prior $p(\boldsymbol{f})$ for the static code $\boldsymbol{f}$ is learned using a Normalizing Flow which consists of 4 Affine Coupling layers [35] each made up of 3 linear layers with hidden size of 256. The final distribution of the NF is set to a standard Gaussian distribution with diagonal covariance matrix.

- **Conditional Normalizing Flow** models the frame latent codes likelihood $p(\boldsymbol{x}_{1:T}|\boldsymbol{f})$. It is composed of 5 Affine coupling layers similar to the ones used for the static prior except for the last layer which is conditioned on the previous frames in order to model the transitions auto-regressively in the same fashion as the AF layer in [108].

The static code is also concatenated to the input of each Affine Coupling layer and thus conditions the transformation. All layers have a hidden size of 256. The transformed latent codes correspond to the dynamic codes $\boldsymbol{\lambda}_{1:T}$ and are of size $n_\lambda = 128$. The final distribution $p(\boldsymbol{\lambda}_{1:T}) = p(\boldsymbol{\lambda}_{1:T}|\boldsymbol{f})$ is parameterized by an LSTM similar to [7]. As uncovered in Ch.3, to achieve a well disentangled model, during training, instead of using the aggregated static code $\boldsymbol{f}$ we additionally make use of the static estimations before aggregation i.e., $\boldsymbol{f}_{1:T}$. We condition the cNF using the randomly shuffled $\boldsymbol{f}_{1:T}$. As such, even if $\boldsymbol{f}_t$ still encompasses dynamic information, due to the random shuffling, that information does not match the frame it conditions and the dynamic information is modeled by the cNF with $\boldsymbol{\lambda}_{1:T}$ which leads to a disentangled model as proved in Sec.A.1.4.

A.3 Additional Results

A.3.1 Complete Tables of Results

Due to the limited space available, some categorical factors accuracies have not been reported into the tables of the main chapter and are reported instead in Tab.A.2. Additionally, in order to also assess if continuous factors are preserved after sampling, we build a regression model based on the frames and then report the R2 score (coefficient of determination) following the same procedure than for the accuracies (Sec.A.2.2). A R2-score of 1. means that the continuous factors have been perfectly preserved after sampling. **A model is disentangled if it achieves a perfect score for all factors.**

In Tab.A.2 it can be observed that, compared to the baseline methods, our proposed model manages to properly disentangle all factors and learn the causal mechanisms. Some competing methods achieve higher scores for some factors, however this can be explained by the decrease in performance regarding the other factors. A model that encodes everything in its dynamic codes will trivially maintain the dynamic when sampling the static code since the latter is uninformative. As an example, for *c-dSprites*, the SPYL method [112] obtains a larger score for the rotation. However, this can be explained by the fact that the shape is wrongly encoded into the dynamic codes and the causal mechanism is discarded by the network;

Table A.2 Disentanglement metrics on *s-dSprites*, *c-dSprites*, *MPI3D* and *LPC-Sprites*, in red and blue are highligted respectively the dynamic factors and the static factors that condition the motion.

	Acc↑	(%)		R^2↑		IS↑	$H(y\|x)$ ↓	$H(y)$ ↑
s-dSprites	color	shape	size	x-pos	y-pos			
DSVAE	79.38	42.10	43.61	0.852	0.920	3.035	0.038	1.17
CDSVAE	96.18	98.22	98.71	0.992	0.989	3.995	0.010	1.37
SPYL	97.76	98.72	98.12	0.998	0.985	3.685	9.5e-3	1.29
Ours	**98.91**	**98.98**	**98.76**	**0.999**	**0.990**	**4.263**	**5.9e-4**	**1.41**

	Acc↑	(%)		R^2↑			IS↑	$H(y\|x)$ ↓	$H(y)$ ↑
c-dSprites	color	shape	size	rot.	x-pos	y-pos			
DSVAE	72.09	32.66	79.38	0.593	0.986	0.985	4.128	0.046	1.41
CDSVAE	96.33	74.32	93.17	0.127	0.986	0.985	4.141	0.028	1.39
SPYL	68.41	33.28	75.10	**0.956**	**0.995**	**0.993**	**4.392**	0.027	1.43
Ours	**98.92**	**98.98**	**98.59**	0.895	0.972	0.972	4.371	**0.008**	**1.44**

	Acc↑	(%)				R^2↑		IS↑	$H(y\|x)$ ↓	$H(y)$ ↑
MPI3D	shape	color	size	camera	light	y-pos	x-pos			
DSVAE	28.29	85.51	65.57	98.81	87.71	0.307	0.312	3.505	0.056	1.19
CDSVAE	43.81	94.09	79.17	**99.99**	**99.94**	0.901	0.707	3.487	0.138	1.26
SPYL	16.78	59.19	50.29	33.32	86.54	**0.994**	**0.977**	3.819	0.043	1.28
Ours	**90.15**	**99.61**	**97.56**	99.97	99.93	0.892	0.878	**3.849**	**0.029**	**1.29**

	Acc↑	(%)					IS↑	$H(y\|x)$ ↓	$H(y)$ ↑
LPCSprites	body	hair	shirt	pants	orient.	action			
CDSVAE	54.94	85.75	99.96	99.54	33.32	67.71	2.737	1.6e-3	1.00
SPYL	70.89	99.89	**100**	**99.74**	33.32	**99.86**	2.735	2.0e-3	**1.01**
Ours	**99.97**	**99.91**	99.99	99.63	**99.94**	99.83	**2.768**	**3.3e-4**	**1.01**

meaning the rotation is trivially preserved after sampling the static code. A similar justification can be made for the *MPI3D* and *LPCSprites* datasets. Overall, the SPYL method seems to be biased towards encoding information in its dynamic codes at the expense of its static code, which might explain why it achieves better performances on *MUG* and *MHAD* when comparing for the actions.

Discussion from the formalism perspective. Using our proposed definition of the factors it is possible to gain further understanding on the results of the baseline methods. These methods enforce simultaneously the extracted dynamic and static codes to be informative and mutually independent. As a result, when the factors are truly independent these constraints are equivalent to our proposed approach and they obtain a disentangled representation, albeit with much more complicated losses e.g. contrastive learning, adversarial training etc. However, by construction, they cannot disentangle when the factors are dependent. The reason is that they generate the sequences by first sampling the codes from a **factorized** prior $p(\boldsymbol{f}, \lambda_{1:T}) = p(\boldsymbol{f})p(\lambda_{1:T})$ and then obtain the frames as $\boldsymbol{x}_t = h(\boldsymbol{f}, \lambda_t)$. Because the transitions between dynamic codes are not conditioned on $\boldsymbol{f}$ and $\boldsymbol{x}_t = h(\boldsymbol{f}, \lambda_t)$, in the best case scenario we can only have $I(\boldsymbol{f}, \lambda_t) = 0 \; \forall t$ and $I(\boldsymbol{f}, \lambda_{1:T}) \geq 0$ meaning information is encoded in both codes and the model is not disentangled. Comparatively our method directly imposes $I(\boldsymbol{f}, \lambda_{1:T}) = 0$ and can thus disentangle the factors.

A.3.2 Shuffle Ablation and Varying Coefficients

In order to confirm what has been demonstrated in Sec.A.1.4, we propose here to compare the performances of the model when ablating the shuffle operation and when varying the β coefficient. The results are provided in Tab.A.3 for the *c-dSprites* dataset.

By looking at Eq.A.58, the β coefficient controls the amount of information enforced in the static code $I(\boldsymbol{f}, \boldsymbol{s})$ and how strongly the static invariance is enforced $KL(q(\boldsymbol{f}_t|\boldsymbol{x}_t)||q(\boldsymbol{f}_t|\boldsymbol{x}_{\pi(t)}))$. On the other hand, the α coefficient controls how much the codes are independent $I(\boldsymbol{\lambda}_{1:T}, \boldsymbol{f}_{\pi_T})$. Intuitively, at fixed α, if the β coefficient is increased (decreased), less (more) information is encouraged in the static code $\boldsymbol{f}$ and the invariance is more (less) enforced resulting in a representation where potentially no (all) information is encoded in $\boldsymbol{f}$ with $I(\boldsymbol{\lambda}_{1:T}, \boldsymbol{f}_{\pi_T})$ trivially satisfied. These results

Table A.3 Disentanglement metrics on *c-dSprites* for varying β and with ablated shuffle operation. In red and blue are highligted respectively the dynamic factors and the static factors that condition the motion. The results boxed are from the model used on all datasets.

c-dSprites		Acc↑ (%)			R^2↑		
		color	shape	size	rot.	x-pos	y-pos
Full model	$\beta = 0.1$	99.92	98.74	99.60	0.716	0.953	0.957
	$\beta = 0.5$	98.92	98.98	98.59	0.895	0.972	0.972
	$\beta = 1.0$	93.92	87.13	91.93	0.966	0.984	0.985
	$\beta = 2.0$	0.08	0.01	0.03	0.998	0.987	0.986
Ablated shuffle		99.94	94.01	96.36	0.0505	0.0561	0.278

are coherent with what can be observed in Tab.A.3 when varying β, confirming Prop.3.3 and Eq.A.58. As stated by Prop.3.3, β needs to be $\ll \alpha$ to avoid all information to go in $\boldsymbol{\lambda}_{1:T}$ but past a certain point it becomes a trade off between having improved performances on the static or dynamic factors. The comparison between our proposal and β-VAE [60] constitutes also a very interesting case. In β-VAE, it is proposed to impose $\beta > \alpha$ to improve disentanglement, at the expense of reconstruction, by more strongly encouraging the posterior to follow a factorized distribution. Conversely, our method does not rely on imposing a factorized distribution for disentanglement but instead makes use directly of the reconstruction term through a cNF. As a result, the disentanglement is instead favored by imposing $\beta < \alpha$ and we do not suffer from the same reconstruction vs disentanglement trade off.

When ablating the shuffle operation, Eq.A.58 is no longer valid. Instead of encouraging the static factors in $\boldsymbol{f}$ ($I(\boldsymbol{f}, \boldsymbol{s})$) we now encourage $\boldsymbol{f}$ to encode everything ($I(\boldsymbol{f}, \boldsymbol{x}_{1:T})$) with no invariance across frames while still enforcing the codes to be independent. As a result the static code $\boldsymbol{f}$ will encode dynamic information and $\boldsymbol{\lambda}_{1:T}$ is encouraged to be non informative. The representation is not properly disentangled and constrained which translates into lower performances for the disentanglement metrics.

A.3.3 Representation Swapping

In Fig.A.2, A.3, A.4 and A.5, we present several qualitative results when swapping the dynamic and static codes for the *MUG*, *LPCSprites*, *MPI3D* and *c-dSprites* datasets. In all examples, clean swaps are obtained and the

static factors are correctly preserved. As discussed in Sec.A.2.2, the motion can only be perfectly preserved if the static factors that causally influence the motion are fixed. For the proposed model, it can be observed that when the conditional static factors are shared between the two test sequences, it results in swapped video clips that correctly preserve the motion. Conversely, when the conditional static factors are different, the swapped video clips exhibit motions that are close to the original dynamic but abide by the causal mechanism. This demonstrates that the model manages to learn the causal mechanism. As an example, for *c-dSprites* in Fig.A.5, when both test sequences share the same shape or size the swapped clip maintains the same rotations and positions respectively. If the objects have different sizes, the swapped clip shows positions that are close to the original driving dynamic code but bounce on the edges according to the new object size.

In Fig.A.6, we also provide a comparison between our proposed method and the competing CDSVAE model when swapping the dynamic and static codes for the *MHAD* dataset. *MHAD* is a much more challenging dataset with more complex motions at higher resolutions. Comparatively, our model achieves cleaner swap motion transfers and provides higher quality non blurry outputs while also better maintaining the identities. As expected however, the results still fall behind what can be achieved by recent human video reenactment SotA methods [138, 140, 139, 15, 170, 23, 163, 24]. Indeed those models rely on high-dimensional structured representations and external human specific signals such as skeletons. In contrast, disentanglement learning aims at obtaining a low-dimensional general representation useful for diverse downstream tasks where the feature dimensions capture the individual semantic concepts, which is different from high dimensional motion transfer tasks. As a result, it cannot preserve detailed spatial information so accurately and general purpose disentanglement of real, in-the-wild, data is still an unsolved problem. As our method manages to properly swap the motion, it shows that the bottleneck is in the decoder and future work may investigates better architectures to reconciliate low dimensional representations with high dimensional reconstructions. This limitation is intrinsic and shared by all the literature in the disentanglement domain [14, 60, 104, 110, 156, 179] and is out of the scope of this work.

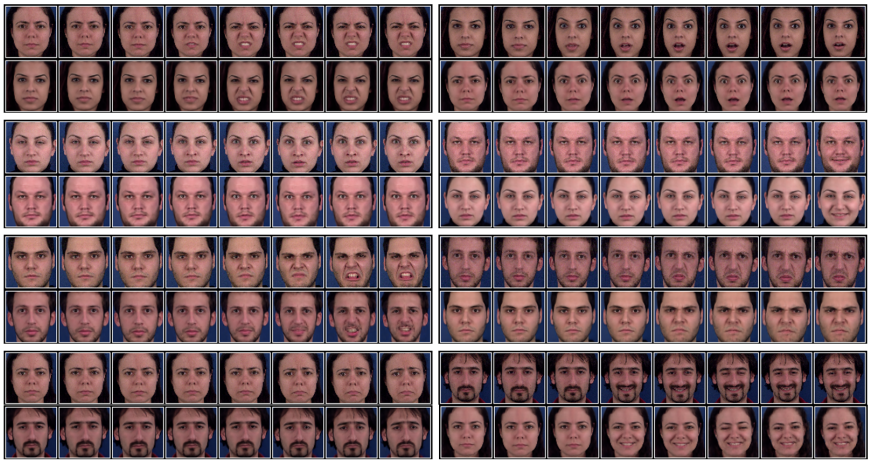


Fig. A.2 Static/dynamic swap results for the MUG dataset. Odd rows are test input sequences while even rows are video clips generated by swapping the static and dynamic codes of the test sequences.

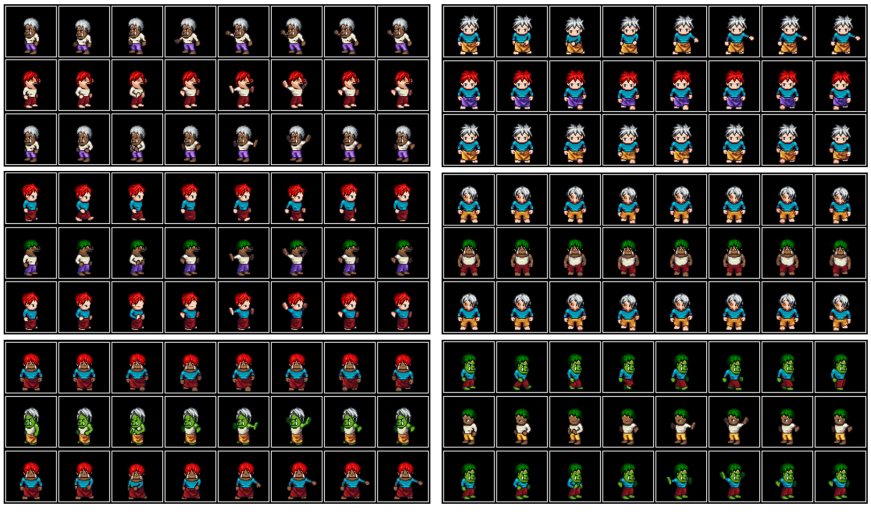


Fig. A.3 Static/dynamic swap results for the LPCSprites dataset. For each group of three sequences, the two first rows are test input sequences while the last row is the video clip generated by swapping the static and dynamic codes of the test sequences.

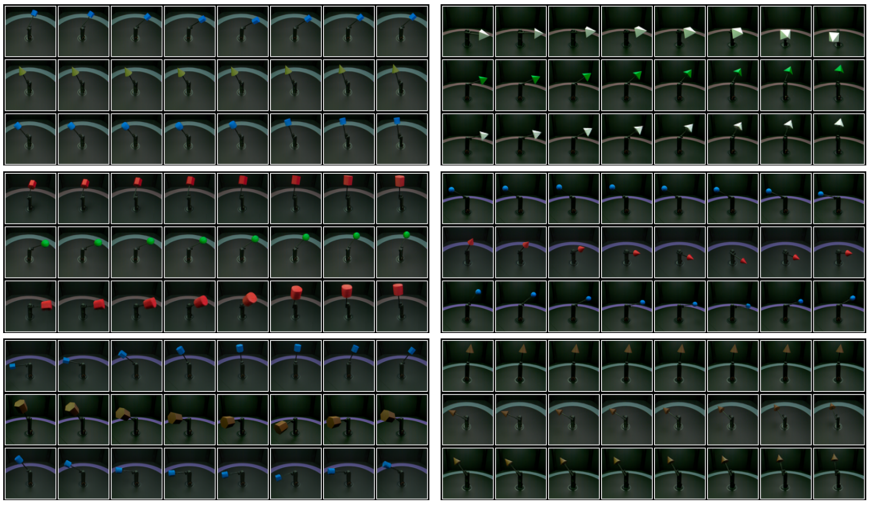


Fig. A.4 Static/dynamic swap results for the MPI3D dataset. For each group of three sequences, the two first rows are test input sequences while the last row is the video clip generated by swapping the static and dynamic codes of the test sequences.

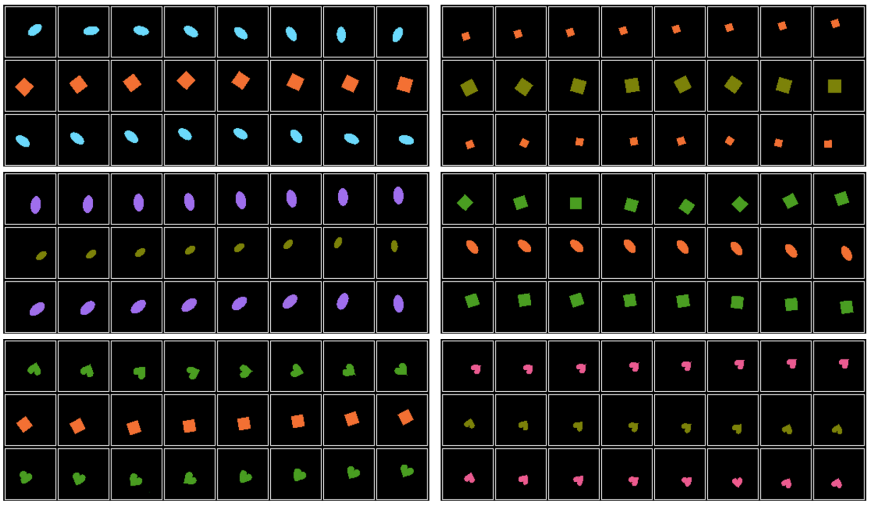


Fig. A.5 Static/dynamic swap results for the c-dSprites dataset. For each group of three sequences, the two first rows are test input sequences while the last row is the video clip generated by swapping the static and dynamic codes of the test sequences.

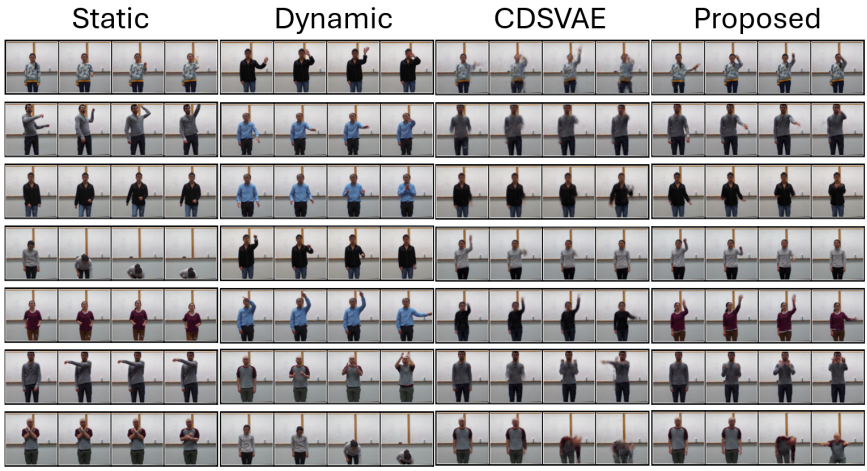


Fig. A.6 Static/dynamic swap results for the MHAD dataset. The first two columns are test input sequences while the last two columns are the video clips generated by swapping the static and dynamic codes of the test sequences using the competing CDSVAE model and our proposed method.

A.3.4 Representation Sampling

In Fig.A.7, A.8 and A.9, we present several qualitative results when sampling either the dynamic or static codes for the *LPCSprites*, *MPI3D* and *c-dSprites* datasets. The discussion is similar to App.A.3.3. On all datasets, the results demonstrate diverse generated factors while preserving the dynamic/static variables following the data causal generative process.

To highlight a potential advantage of the static/dynamic disentanglement for unconditional generation, we also compare in Fig.A.10 our proposed model with the same model but without a global static code (i.e., no disentanglement). It can be observed that by using a global static code for generating new sequences, the aspects of the shapes are better preserved across time steps. The colors, shapes and sizes are provided to the conditional Normalizing Flow via $\boldsymbol{f}$ to reconstruct each frame allowing to better maintain these factors and avoid flickering. This implies that adding a global code could help video generation problems in better maintaining the temporal coherence.

A.3.5 Interpolation in Latent Space

In Fig.A.11 and A.12, we show results when interpolating between sequences in both the static and dynamic latent spaces. These figures demonstrate that the learned latent spaces are smooth and meaningful. In all examples, we can observe a smooth transition between the interpolated factors while the fixed factors are maintained, further demonstrating proper disentanglement.

A.3.6 Latent embedding visualization

In order to visualize the latent spaces, we plot in Fig.A.13 the 2D embedding of the static and dynamic codes, computed using t-SNE [148] on all *s-dSprites* test samples. For both the static and dynamic codes, we color the embedded points based on the static factors $\boldsymbol{s}$ i.e., the objects colors, shapes and sizes. We can observe that the static code $\boldsymbol{f}$ has a clear structure with clusters that correspond to all unique combinations of static factors. Conversely, the dynamic codes do not show any structure with the static factors and appear independent to $\boldsymbol{s}$. In Fig.A.14, the same experiment is repeated but with the embedded points colored based on the dynamic factors $\boldsymbol{d}_{1:T}$ i.e., the positions of the moving shapes and it is now only the dynamic codes that display a clear structure with respect to the factors. We thus conclude that the method subspaces are disentangled.

A.4 Details on the Limitations

In Sec.3.5 of the main chapter, several limitations have been highlighted. These limits are not specific to our proposed model and are intrinsic to the two-factor static/dynamic disentanglement approach. Following the problem formulation in Sec.3.3, the static variables are defined as invariant in each sequence while the dynamic variables are defined as varying almost surely in the dataset (but not necessarily in every sequence). This has two consequences:

1. The static factors only capture the factors that are invariant in **all** sequences.
2. The static/dynamic factors are defined by the dataset itself.

For the first consequence, it means that by having a single static code f, the network will only ever be able to capture in that code at maximum the factors that are static in all sequences. A factor static for a specific sequence but that sometimes may vary in others will be encoded as a dynamic factor. The factors are 'labeled' by the network as static or dynamic at the dataset level and not at the sequence level. As an illustrative example, in Fig.A.15 for *c-dSprites*, we take sequences with a fixed rotation and swap the dynamic codes. The resulting sequences have different rotations, meaning the shape angle is encoded in the dynamic code even if it was static for the specific sequence of origin.

For the second consequence, while it might be beneficial to have the network uncovering the factors and the causal relationships without supervision, it means that if one wants to specify a particular factor as dynamic instead of static then it restricts the datasets that can be used. As an illustrative example, in Fig.A.15, the network is trained with *LPCSprites* sequences where the cartoon characters are turning and only then the orientation of the characters is encoded as a dynamic factor. Overall it means the network **discovers** the static/dynamic factors **depending on the dataset**. This has also other major implications. For evaluation, the common benchmark involves the use of labels associated with the videos [99, 7, 180]. However, the notion of static/dynamic captured is intrinsically tied to the dataset and not to the labels. These labels are used for the metrics but might not provide the full picture of the possible dependencies and motions that are present in the dataset. This might explain the gap observed for the quantitative results when assessing on real-world datasets where the relation between the ground-truth generation process and labels is not tightly controlled unlike for the synthetic datasets. This highlights the need for future work to provide novel more robust evaluation processes. Finally, it also shows that altering the dataset e.g., with data augmentation like CDSVAE [7], actually defines what is static/dynamic, the augmentations affect the dataset and thus change what will be captured. Instead of disentangling the satic/dynamic of the data, the model will disentangle the augmentations. Similarly,using supervised learning with labels will define the disentanglement instead of discovering the true factors.

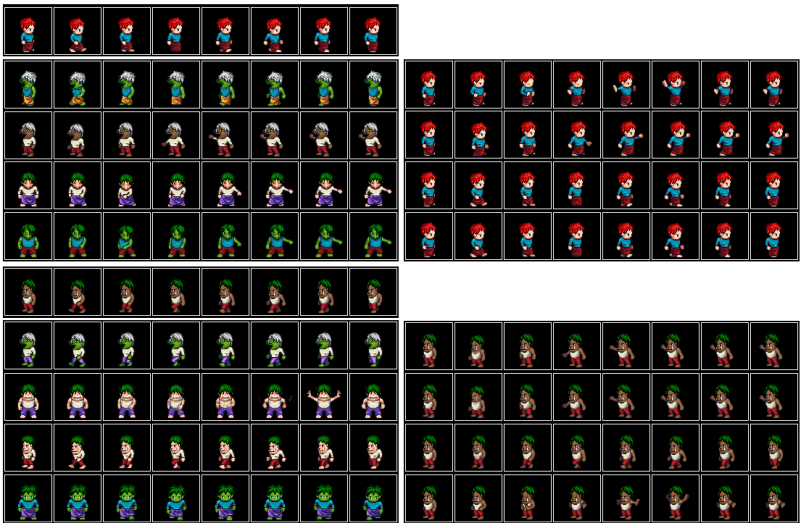


Fig. A.7 Static/Dynamic generation results for the LPCSprites dataset. Row 1 and 6 are test sequences. For the other rows, the left (right) column comprises video clips generated by fixing the dynamic (static) codes to the value given by the test sequence and sampling the static (dynamic) variables from the prior.

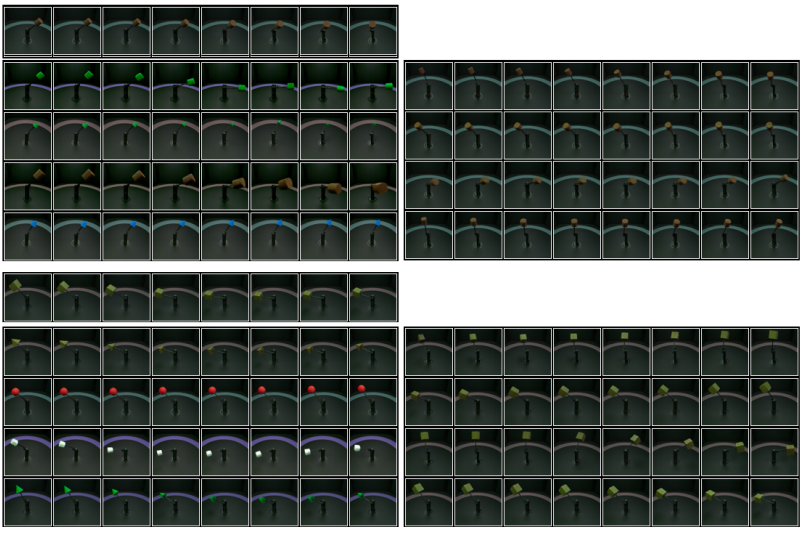


Fig. A.8 Static/Dynamic generation results for the MPI3D dataset. Row 1 and 6 are test sequences. For the other rows, the left (right) column comprises video clips generated by fixing the dynamic (static) codes to the value given by the test sequence and sampling the static (dynamic) variables from the prior.

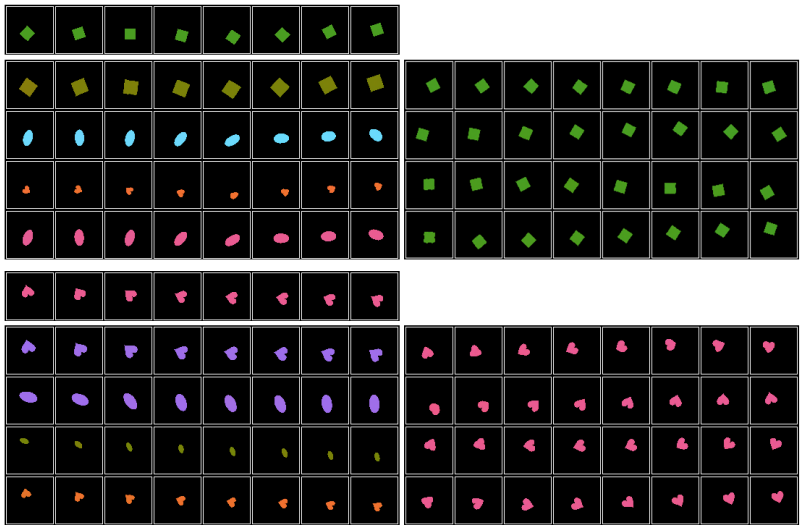


Fig. A.9 Static/Dynamic generation results for the c-dSprites dataset. Row 1 and 6 are test sequences. For the other rows, the left (right) column comprises video clips generated by fixing the dynamic (static) codes to the value given by the test sequence and sampling the static (dynamic) variables from the prior.

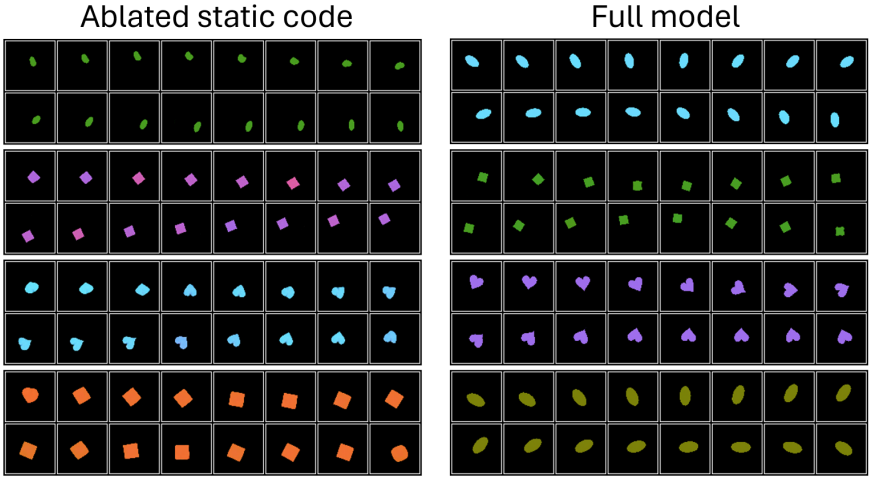


Fig. A.10 Unconditional generation results for the c-dSprites dataset. On the left are presented samples from the model without a global static code and on the right are shown results for the full model. Sequences generated with a global static code better maintain the shapes aspect across time steps.

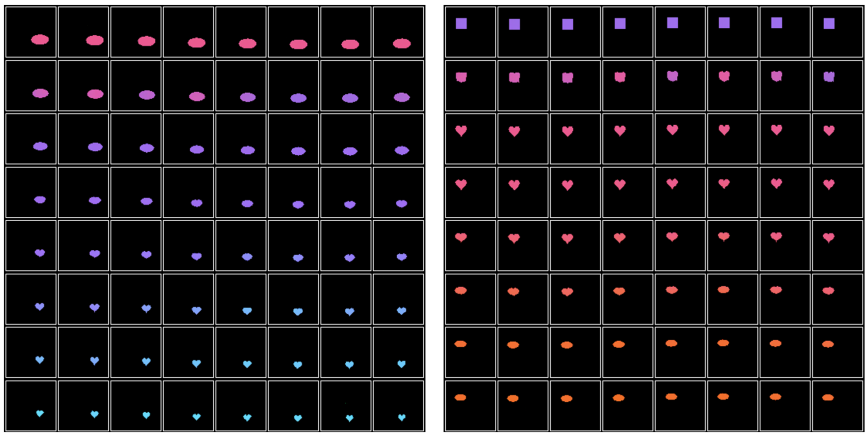


Fig. A.11 Interpolation in the latent space for s-dSprites. In each of the two columns, the first and last rows share the same dynamic codes $\lambda_{1:T}$ but have different static codes f. The rows in between are generated by fixing the dynamic codes and interpolating the static one using cosine interpolation.

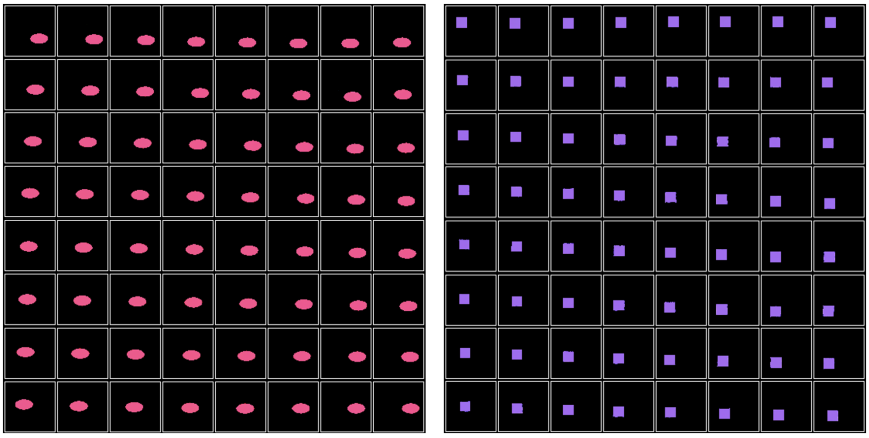


Fig. A.12 Interpolation in the latent space for s-dSprites. In each of the two columns, the first and last rows share the same static code f, but have different dynamic codes $\lambda_{1:T}$. The rows in between are generated by fixing the static code and interpolating the dynamic ones using cosine interpolation.

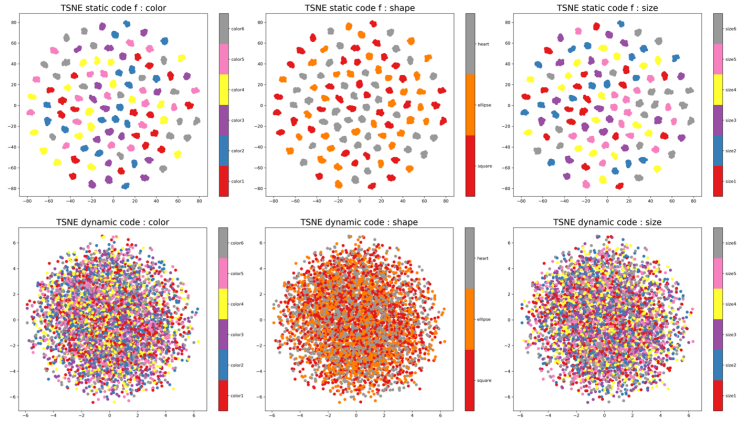


Fig. A.13 2D embedding of the static (top row) and dynamic (bottom row) codes computed using t-SNE on all *s-dSprites* test samples. In each row we display the embedding with points colored according to the different static factors of the dataset. The static code embedding perfectly clusters all static factors combinations while the dynamic codes appear invariant to the static variables.

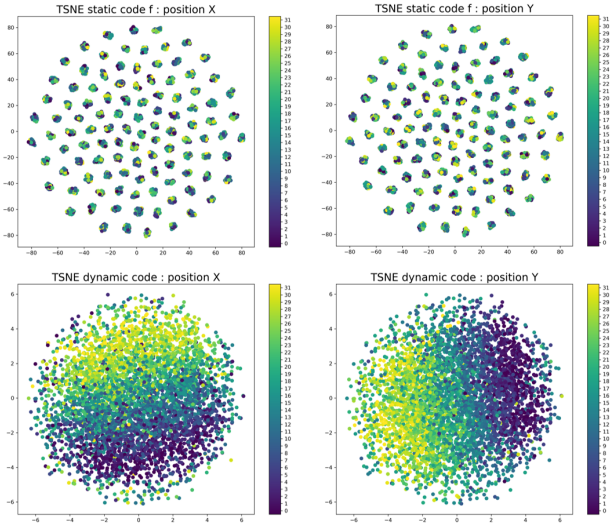


Fig. A.14 2D embedding of the static (top row) and dynamic (bottom row) codes computed using t-SNE on all *s-dSprites* test samples. In each row we display the embedding with points colored according to the different dynamic factors of the dataset i.e., the 32 possible positions of the shapes. The dynamic code embedding displays a structure with respect to the dynamic factors while the static codes appear invariant to the dynamic variables.

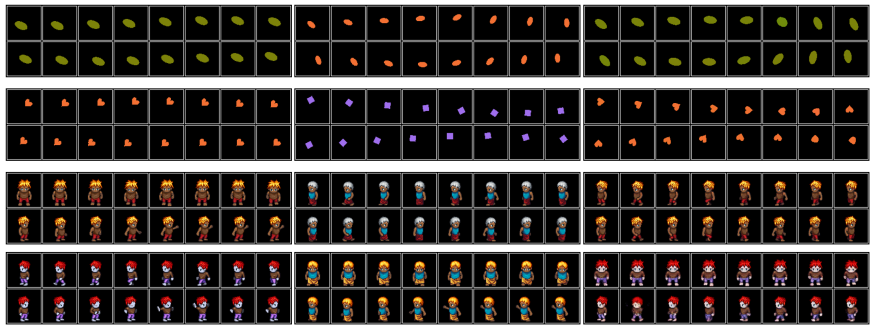


Fig. A.15 Illustration of the limitations of the two-factor static/dynamic disentanglement. The left and middle columns are test input sequences and the right column are video clips generated by swapping the static and dynamic codes of the test sequences.

B

Supplementary Material for Chapter 4

B.1 Proofs and Detailed Derivations

This appendix provides the formal proofs of the main theoretical results stated in Sec.4.4. We begin by introducing auxiliary results that will be used throughout the proofs. We then establish identifiability under the independence assumption (Prop.4.1) and discuss its connection to classical Independent Component Analysis (ICA). Finally, we present the proof of identifiability under dependent latent factors (Prop.4.2) and highlight the key differences with respect to the independent case.

Throughout this appendix, as specified in Def.4.1, all functions are assumed to be sufficiently smooth, and all mappings are assumed to be invertible on their respective domains unless stated otherwise.

B.1.1 Preliminary Results

Preliminary. *Let $\psi : [0,1]^d \to [0,1]^d$ be a smooth and invertible diffeomorphism that maps the hypercube onto itself and sends boundary points to boundary*

points. for every $u \in [0,1]^d$, the Jacobian of ψ admits a factorization

$$J_\psi(u) = D_1(u)\, R(u)\, D_2(u), \tag{B.1}$$

where $D_1(u)$ and $D_2(u)$ are invertible diagonal matrices and $R(u) \in O(d)$. Additionally, the Jacobian field J_ψ satisfies the integrability condition corresponding to a globally defined smooth mapping. Then there exists a signed permutation matrix P and smooth, strictly monotone functions $\varphi_i : [0,1] \to [0,1]$ such that

$$\psi(u_1, \ldots, u_d) = P\big(\varphi_1(u_1), \ldots, \varphi_d(u_d)\big). \tag{B.2}$$

In particular, up to a fixed signed permutation of coordinates, ψ is a global coordinate-wise (unitwise) reparametrization of the hypercube.

Proof. By construction, the Jacobian of ψ admits, at every point $u \in [0,1]^d$, the factorized form

$$J_\psi(u) = D_1(u)\, R(u)\, D_2(u), \tag{B.3}$$

where $D_1(u)$ and $D_2(u)$ are invertible diagonal matrices, and $R(u) \in O(d)$ is an orthogonal matrix. This decomposition holds smoothly on the entire domain, including its boundary. In particular, for each boundary point $u \in \partial[0,1]^d$, there exists a neighborhood on which the Jacobian preserves this diagonal–orthogonal–diagonal structure.

Let $F_i^\pm$ denote the $(d-1)$-dimensional faces of the hypercube defined by

$$F_i^- = \{u \in [0,1]^d \mid u_i = 0\}, \qquad F_i^+ = \{u \in [0,1]^d \mid u_i = 1\}. \tag{B.4}$$

For any point $p \in F_i^\pm$, the image $\psi(p)$ lies on the boundary $\partial[0,1]^d$, and hence belongs to some face $F_k^\pm$ for a certain index $k \in \{1, \ldots, d\}$. The tangent space at p to the face $F_i^\pm$ is given by

$$T_p F_i^\pm = \operatorname{span}\{e_j \mid j \neq i\}, \tag{B.5}$$

where $\{e_1, \ldots, e_d\}$ is the canonical basis of $\mathbb{R}^d$. Since ψ maps $\partial[0,1]^d$ onto itself, its differential must map tangent directions along a face to tangent directions of the image face. Equivalently, the differential maps the normal vector to a face to a vector normal to the image face, up to a nonzero scalar factor.

Let $\boldsymbol{n}_i = \pm e_i$ denote the unit normal vector to $F_i^{\pm}$ at p. Then

$$J_\psi(p)\,\boldsymbol{n}_i = D_1(p)\,R(p)\,D_2(p)\,\boldsymbol{n}_i = \lambda_i\,D_1(p)\,R(p)\,e_i, \tag{B.6}$$

for some nonzero scalar λ_i corresponding to the ith diagonal entry of $D_2(p)$. Because the image of $\boldsymbol{n}_i$ must be normal to the image face $F_k^{\pm}$, the vector $R(p)e_i$ must be colinear with $\pm e_k$. As $R(p)$ is orthogonal, this implies that its (k,i)-entry satisfies $R_{ki}(p) = \pm 1$, and all other elements in the ith column and kth row of $R(p)$ vanish. Hence, locally, $R(p)$ acts as a signed permutation matrix exchanging coordinates i and k (possibly flipping their orientation).

Since ψ (and therefore $R(u)$) is continuous, and the set of signed permutation matrices is discrete, the local signed permutation structure of $R(u)$ must remain constant on each connected component of the boundary where it is defined. As each face $F_i^{\pm}$ of the hypercube is connected, this implies that ψ maps each face $F_i^{\pm}$ diffeomorphically onto a single face $F_k^{\pm}$, with the correspondence between indices (i,k) determined by a fixed signed permutation.

To simplify the notation, we now restrict our attention to the three-dimensional case ($d = 3$); the argument extends verbatim to arbitrary dimensions. Moreover, without loss of generality, we assume that the signed permutation describing the correspondence between boundary faces is the identity. This assumption merely fixes the coordinate labeling and does not affect the generality of the proof, since any other signed permutation can be factored out and reintroduced at the end of the argument.

Let F_1, F_2, and F_3 denote the three faces of the cube $[0,1]^3$ intersecting at the origin, so that each face F_i is orthogonal to the coordinate axis e_i, with outward normal vector $\boldsymbol{n}_i = -e_i$. The pairwise intersections of these faces correspond to the coordinate axes, and their triple intersection corresponds to the origin.

As established in the previous section (Eq.B.6), the Jacobian matrix of ψ restricted to each face must preserve the tangent directions of that face. For $u \in F_1$, F_2, or F_3, the Jacobian therefore admits the following face-specific structures:

$$J_{\psi|F_1}(u) = \begin{pmatrix} d_{1,1}(u) & 0 & 0 \\ 0 & d_{1,2}(u) & 0 \\ 0 & 0 & d_{1,3}(u) \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & a & b \\ 0 & c & d \end{pmatrix} \begin{pmatrix} d_{2,1}(u) & 0 & 0 \\ 0 & d_{2,2}(u) & 0 \\ 0 & 0 & d_{2,3}(u) \end{pmatrix}, \tag{B.7}$$

$$J_{\psi|F_2}(u) = \begin{pmatrix} d_{1,1}(u) & 0 & 0 \\ 0 & d_{1,2}(u) & 0 \\ 0 & 0 & d_{1,3}(u) \end{pmatrix} \begin{pmatrix} a & 0 & b \\ 0 & 1 & 0 \\ c & 0 & d \end{pmatrix} \begin{pmatrix} d_{2,1}(u) & 0 & 0 \\ 0 & d_{2,2}(u) & 0 \\ 0 & 0 & d_{2,3}(u) \end{pmatrix}, \tag{B.8}$$

$$J_{\psi|F_3}(u) = \begin{pmatrix} d_{1,1}(u) & 0 & 0 \\ 0 & d_{1,2}(u) & 0 \\ 0 & 0 & d_{1,3}(u) \end{pmatrix} \begin{pmatrix} a & b & 0 \\ c & d & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} d_{2,1}(u) & 0 & 0 \\ 0 & d_{2,2}(u) & 0 \\ 0 & 0 & d_{2,3}(u) \end{pmatrix}. \tag{B.9}$$

At the pairwise intersections of these faces (that is, along the coordinate axes) and at their triple intersection (the origin), the orthogonal component necessarily satisfies $R(u) = I_3$. Consequently, the Jacobian at these intersection points simplifies to a purely diagonal form:

$$J_{\psi|\text{intersection}}(u) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} d_1^*(u) & 0 & 0 \\ 0 & d_2^*(u) & 0 \\ 0 & 0 & d_3^*(u) \end{pmatrix}, \tag{B.10}$$

where $d_i^*(u)$ denote the diagonal entries of the product of $D_1(u)$ and $D_2(u)$. If the global signed permutation of faces were not the identity, the same conclusion would hold up to a fixed permutation and possible sign changes of these diagonal components.

While we must validate the orthogonal shape of the jacobian constraint everywhere, there is also another constraint that must be validated by the mapping. For ψ to be globally realizable as a smooth diffeomorphism, the Jacobian field $J_\psi(u)$ must satisfy the standard *integrability condition*, which ensures that it corresponds to the gradient of a smooth mapping. Specifically, for each $i, j, k \in \{1, 2, 3\}$, the mixed partial derivatives must commute:

$$\frac{\partial^2 \psi_i}{\partial u_j \, \partial u_k} = \frac{\partial^2 \psi_i}{\partial u_k \, \partial u_j}. \tag{B.11}$$

When the second partial derivatives are continuous, this follows from the theorem of Schwarz, which implies that the Hessian matrices of the coordinate functions of ψ are symmetric. In differential form, this condition can be equivalently expressed as

$$\frac{\partial J_\psi}{\partial u_j} e_k = \frac{\partial J_\psi}{\partial u_k} e_j, \tag{B.12}$$

ensuring that the curl of each row of J_ψ vanishes.

This integrability constraint imposes compatibility relations among the entries of $D_1(u)$, $D_2(u)$, and $R(u)$, ensuring that the Jacobian structure de-

fined above is not only pointwise valid but also arises from a globally consistent smooth mapping ψ.

At the corner of the cube, where all three faces intersect, we have already established that $R(u) = I_3$. Furthermore, since $R(u)$ remains the identity along each pairwise intersection of faces, its value is fixed and constant in all directions emanating from the corner. Applying the integrability condition in this setting thus constrains the diagonal components of the Jacobian. Writing

$$J_{\psi|\text{corner}}(u) = \begin{pmatrix} d_1^*(u) & 0 & 0 \\ 0 & d_2^*(u) & 0 \\ 0 & 0 & d_3^*(u) \end{pmatrix}, \tag{B.13}$$

the mixed partial symmetry condition $\partial_{u_i}\partial_{u_j}\psi_k = \partial_{u_j}\partial_{u_i}\psi_k$ implies that cross-derivatives of distinct coordinates must vanish. Consequently, each diagonal entry $d_i^*(u)$ can depend only on its corresponding coordinate u_i, yielding

$$\begin{pmatrix} d_1^*(u) & 0 & 0 \\ 0 & d_2^*(u) & 0 \\ 0 & 0 & d_3^*(u) \end{pmatrix} = \begin{pmatrix} d_1^*(u_1) & 0 & 0 \\ 0 & d_2^*(u_2) & 0 \\ 0 & 0 & d_3^*(u_3) \end{pmatrix}. \tag{B.14}$$

Hence, near the corner, ψ reduces to a *unitwise reparametrization* of the coordinates, meaning that each coordinate function ψ_i depends solely on its corresponding input variable u_i. Showing how the integrability condition affects the possible mapping at the corners of the hypercube and thus the importance of this condition.

We now apply the integrability condition to the restriction of ψ on a boundary face, for instance $F_1 = \{u_1 = 0\}$. On this face, the Jacobian admits the form

$$J_{\psi|F_1}(u) = \begin{pmatrix} d_{1,1}(u) & 0 & 0 \\ 0 & d_{1,2}(u) & 0 \\ 0 & 0 & d_{1,3}(u) \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & a & b \\ 0 & c & d \end{pmatrix} \begin{pmatrix} d_{2,1}(u) & 0 & 0 \\ 0 & d_{2,2}(u) & 0 \\ 0 & 0 & d_{2,3}(u) \end{pmatrix}, \tag{B.15}$$

where $R(u) = \left(\begin{smallmatrix} 1 & 0 & 0 \\ 0 & a & b \\ 0 & c & d \end{smallmatrix}\right)$ acts as the local orthogonal component on F_1. By construction, $R(u)$ fixes both its first column and its first row across the entire face, reflecting the fact that the normal direction e_1 is invariant under ψ along F_1.

Since $R(u)$ is constant with respect to u_2 and u_3 in its first column and row, differentiating $J_{\psi|F_1}$ along the tangent directions u_2 and u_3 does not

alter the components corresponding to e_1. The integrability condition,

$$\frac{\partial J_\psi}{\partial u_i} e_j = \frac{\partial J_\psi}{\partial u_j} e_i, \qquad \forall\, i,j \in \{1,2,3\}, \tag{B.16}$$

then enforces that the mixed derivatives involving u_1 and any tangent direction to F_1 vanish in the corresponding entries of J_ψ. In particular, the entries associated with the tangent scaling factors $(d_{1,2}, d_{1,3}, d_{2,2}, d_{2,3})$ must satisfy

$$\frac{\partial d_{1,2}}{\partial u_1} = \frac{\partial d_{1,3}}{\partial u_1} = \frac{\partial d_{2,2}}{\partial u_1} = \frac{\partial d_{2,3}}{\partial u_1} = 0. \tag{B.17}$$

Consequently, these quantities cannot depend on u_1, i.e.

$$d_{1,2}(u) = d_{1,2}(u_2, u_3), \quad d_{1,3}(u) = d_{1,3}(u_2, u_3), \tag{B.18}$$

$$d_{2,2}(u) = d_{2,2}(u_2, u_3), \quad d_{2,3}(u) = d_{2,3}(u_2, u_3). \tag{B.19}$$

Geometrically, this expresses the fact that along the face F_1, variations parallel to the face (in u_2 and u_3) occur independently of the normal coordinate u_1, since the orthogonal component $R(u)$ preserves the orientation of the normal direction. The integrability constraint thus enforces that the scaling along tangent directions to a face cannot depend on the coordinate normal to that face.

The foregoing discussion applies mutatis mutandis to every face, so that analogous conclusions hold for F_2 and F_3. In particular, on the pairwise intersection $F_1 \cap F_3$ we have (recall that $R = I$ on intersections)

$$J_{\psi|F_1\cap F_3}(u) = I_3 \begin{pmatrix} d_1^*(u) & 0 & 0 \\ 0 & d_2^*(u) & 0 \\ 0 & 0 & d_3^*(u) \end{pmatrix} = I_3 \begin{pmatrix} d_1^*(u_1,u_2) & 0 & 0 \\ 0 & d_2^*(u_2) & 0 \\ 0 & 0 & d_3^*(u_2,u_3) \end{pmatrix}, \tag{B.20}$$

where the displayed functional dependencies record the constraints obtained from applying integrability on the adjacent faces. Thus, when one moves along the intersection curve (which is parametrized by u_2), the mapping acts as a unitwise reparametrization in the u_2–direction; perpendicular variations may still modify the scalings d_1^* and d_3^* through their explicit dependence on u_2. At the triple intersection (the corner) this specialization reduces to the purely unitwise reparametrization obtained earlier, which is coherent with what we obtained earlier.

Consider now, at that intersection $F_1 \cap F_3$, parameterized by the coordinate u_2, along which the Jacobian takes the simplified form (B.20), the

integrability condition in the e_3-direction on this one-dimensional curve, namely

$$\frac{\partial J_\psi}{\partial u_2} e_3 = \frac{\partial J_\psi}{\partial u_3} e_2. \tag{B.21}$$

Because on $F_1 \cap F_3$ we have $R(u) = I_3$ (by construction) and $J_\psi(u) = D^*(u)$ is diagonal, we can explicitly compute for the left-hand side,

$$\frac{\partial J_\psi}{\partial u_2} e_3 = \frac{\partial}{\partial u_2} \begin{pmatrix} 0 \\ 0 \\ d_3^*(u_2, u_3) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \dfrac{\partial d_3^*(u_2, u_3)}{\partial u_2} \end{pmatrix}. \tag{B.22}$$

For the right-hand side, differentiation is taken along the u_3-direction, which is tangent to the face F_1. In this case, $R(u)$ retains its first row and column fixed (as established earlier), but the remaining entries of $R(u)$ may vary smoothly with u_3. Thus we obtain

$$\frac{\partial J_\psi}{\partial u_3} e_2 = \frac{\partial}{\partial u_3} \big(R(u) D^*(u)\big) e_2 = \frac{\partial R(u)}{\partial u_3} D^*(u) e_2 + \frac{\partial D^*(u)}{\partial u_3} e_2. \tag{B.23}$$

The second term in (B.23) vanishes when evaluated on e_2 (because $d_2^*(u)$ depends only on u_2 along the intersection). Hence, we are left with

$$\frac{\partial J_\psi}{\partial u_3} e_2 = \frac{\partial R(u)}{\partial u_3} d_2^*(u_2) e_2, \tag{B.24}$$

Since $R(u)$ fixes e_1 and its first row/column and is fixed in u_2, $\frac{\partial R(u)}{\partial u_3}$ is a smooth matrix-valued function depending only on u_3 along the intersection.

Equating the two sides of (B.21) then yields the scalar constraint

$$\frac{\partial d_3^*(u_2, u_3)}{\partial u_2} = \xi(u_3)\, d_2^*(u_2), \tag{B.25}$$

where $\xi(u_3)$ is a scalar function determined by the infinitesimal rotation of $R(u)$ in the (e_2, e_3)-plane. Eq.B.25 expresses that d_3^* must vary with u_2 according to a first-order linear relation whose coefficient depends only on u_3.

Finally, recall that the intersection $F_1 \cap F_3$ contains the corner point where ψ reduces to a unitwise reparametrization. At this corner, the func-

tion d_3^* must coincide with the corresponding unitwise scaling, which enforces $\frac{\partial d_3^*(u_2,u_3)}{\partial u_2} = 0$. However, $d_2^*(u_2) \neq 0$ as the function is smooth and invertible everywhere. Hence, $\xi(u_3) = 0$ at the corner and thus also on the whole intersection $F_1 \cap F_3$ parametrized by u_2. As the same proof can be applied to d_1^*, overall we obtain that

$$J_{\psi|F_1\cap F_3}(u) = I_3 \begin{pmatrix} d_1^*(u_1) & 0 & 0 \\ 0 & d_2^*(u_2) & 0 \\ 0 & 0 & d_3^*(u_3) \end{pmatrix}. \tag{B.26}$$

This shows that along $F_1 \cap F_3$, the mapping remains a unitwise reparametrization: each coordinate evolves independently. This conclusion is consistent with the earlier analysis at the corner and extends the unitwise reparametrization property to all intersection curves between adjacent faces F_1, F_2, F_3.

Eq.B.20 established that on the intersections of adjacent faces, the Jacobian reduces to a diagonal form, corresponding to a coordinate-wise (unitwise) reparametrization, possibly composed with a fixed signed permutation. However, this constraint initially applies only along the intersection curves themselves; a priori, when moving in a direction perpendicular to such a curve—i.e., transversally to the intersection of two faces—the mapping need not remain a pure reparametrization. The refined condition (B.26) strengthens this conclusion. Indeed, along the intersection $F_1 \cap F_3$, the integrability condition has shown that not only along the curve (parameterized by u_2) but also infinitesimally in directions orthogonal to it, the orthogonal component $R(u)$ remains fixed and equal to I_3. This means that the differential of ψ is purely diagonal in a full neighborhood of each intersection curve:

$$J_\psi(u) = I_3 \begin{pmatrix} d_1(u_1) & 0 & 0 \\ 0 & d_2(u_2) & 0 \\ 0 & 0 & d_3(u_3) \end{pmatrix} \quad \text{for } u \text{ in a neighborhood of } F_i \cap F_j. \tag{B.27}$$

In such a neighborhood, infinitesimal displacements along coordinate directions remain aligned with the coordinate axes. In particular, straight coordinate lines on the boundary, defined by fixing all but one coordinate, are mapped by ψ to straight coordinate lines in the image, with no curvature introduced.

Now, consider moving an infinitesimal step du away from the intersection curve into the interior of the cube. Because $R(u)$ is continuous,

$R(u) = I_3$ along the intersection, and the set of orthogonal matrices is discrete in a neighborhood of the identity under the signed permutation constraint, the orthogonal component must remain constant in a connected neighborhood:

$$R(u) = I_3 \qquad \text{for all } u \text{ in a connected neighborhood of } F_i \cap F_j. \tag{B.28}$$

Hence, within this neighborhood, ψ continues to act as a unitwise reparameterization:

$$J_\psi(u) = \operatorname{diag}\big(d_1(u_1),\, d_2(u_2),\, d_3(u_3)\big). \tag{B.29}$$

Starting from one intersection curve (say $F_1 \cap F_3$), we can apply the same argument to adjacent intersection curves $F_1 \cap F_2$ and $F_2 \cap F_3$. Each of these intersections is connected and shares a corner point where $R(u) = I_3$, ensuring compatibility across overlaps. By continuity and connectedness of the cube $[0,1]^3$, these local neighborhoods form a connected cover of the entire domain. Consequently, the property (B.29) extends globally throughout $[0,1]^3$.

Therefore, on the whole cube, the Jacobian of ψ has a globally diagonal structure with each diagonal element depending only on its corresponding coordinate:

$$J_\psi(u) = \begin{pmatrix} d_1(u_1) & 0 & 0 \\ 0 & d_2(u_2) & 0 \\ 0 & 0 & d_3(u_3) \end{pmatrix}. \tag{B.30}$$

Integrating this expression along each coordinate direction yields

$$\psi(u_1, u_2, u_3) = \big(\varphi_1(u_1),\, \varphi_2(u_2),\, \varphi_3(u_3)\big), \tag{B.31}$$

where each $\varphi_i : [0,1] \to [0,1]$ is a smooth, strictly monotone reparametrization function.

In summary, the integrability and continuity conditions jointly ensure that the local unitwise reparametrization property propagates from intersections of faces to their neighborhoods and ultimately to the entire cube. Thus, the diffeomorphism ψ on $[0,1]^3$ must be a global coordinate-wise reparametrization (up to a fixed signed permutation, if present), concluding the proof. □

B.1.2 Proof of Proposition 4.1

Proposition 4.1. *Let the observed data $\boldsymbol{x} \in \mathbb{R}^n$ lie on a d-dimensional manifold generated by latent variables $\boldsymbol{z} \in \mathbb{R}^d$, with $d \ll n$, through a smooth and invertible mapping $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^n$ satisfying Asm.4.1. Each latent coordinate z_i corresponds to a distinct underlying concept. Suppose further that the latent variables are statistically independent, $p(\boldsymbol{z}) = \prod_i p(z_i)$, and that at most one component z_i follows a Gaussian distribution. Then, the generative process $(\boldsymbol{f}, p_{\boldsymbol{z}})$ is identifiable in the sense of Def.4.1.*

Proof. Assume $\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^n$ is a smooth diffeomorphism satisfying Asm.4.1 (the columns of its Jacobian are orthogonal) and that

$$\boldsymbol{f}_* p_{\boldsymbol{z}} = \boldsymbol{g}_* p_{\hat{\boldsymbol{z}}} \tag{B.32}$$

where $p_{\hat{\boldsymbol{z}}} = \prod_{i=1}^d p_{\hat{z}_i}$ i.e., the latent coordinates $\hat{\boldsymbol{z}} \in \mathbb{R}^d$ are independent. We must show that $\boldsymbol{g}$ recovers $\boldsymbol{f}$ up to the ambiguities allowed in Def.4.1 (a permutation and element-wise reparameterizations).

Because $\boldsymbol{f}$ and $\boldsymbol{g}$ are diffeomorphisms onto the same d-dimensional manifold, define

$$\boldsymbol{h} := \boldsymbol{f}^{-1} \circ \boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^d. \tag{B.33}$$

By the pushforward identity we have

$$\boldsymbol{h}_* p_{\hat{\boldsymbol{z}}} = p_{\boldsymbol{z}}. \tag{B.34}$$

Thus $\boldsymbol{h}$ is a smooth bijection that transforms the product density $p_{\hat{\boldsymbol{z}}}$ into the product density $p_{\boldsymbol{z}}$.

Fix $u \in \mathbb{R}^d$ and let $v = \boldsymbol{h}(u)$. By Asm.4.1, at the corresponding points the Jacobians of $\boldsymbol{f}$ and $\boldsymbol{g}$ admit the column-orthogonal factorization

$$J_{\boldsymbol{f}}(v) = Q_{\boldsymbol{f}}(v) D_{\boldsymbol{f}}(v), \qquad J_{\boldsymbol{g}}(u) = Q_{\boldsymbol{g}}(u) D_{\boldsymbol{g}}(u), \tag{B.35}$$

where for each argument $Q_{\boldsymbol{f}}(v), Q_{\boldsymbol{g}}(u) \in \mathbb{R}^{n \times d}$ have orthonormal columns ($Q^T Q = I_d$), and $D_{\boldsymbol{f}}(v), D_{\boldsymbol{g}}(u) \in \mathbb{R}^{d \times d}$ are diagonal matrices. Because $\boldsymbol{f}$ is a local diffeomorphism, the derivative of $\boldsymbol{f}^{-1}$ at $x = \boldsymbol{f}(v)$ equals the left inverse of $J_{\boldsymbol{f}}(v)$:

$$D(\boldsymbol{f}^{-1})(x) = (J_{\boldsymbol{f}}(v))^+ = (J_{\boldsymbol{f}}(v)^T J_{\boldsymbol{f}}(v))^{-1} J_{\boldsymbol{f}}(v)^T = D_{\boldsymbol{f}}(v)^{-1} Q_{\boldsymbol{f}}(v)^T \tag{B.36}$$

where we used $J_{\boldsymbol{f}} = Q_{\boldsymbol{f}} D_{\boldsymbol{f}}$ and $Q_{\boldsymbol{f}}^T Q_{\boldsymbol{f}} = I_d$. Consequently, the Jacobian of

$\boldsymbol{h}$ at u is

$$J_{\boldsymbol{h}}(u) = D(\boldsymbol{f}^{-1})(\boldsymbol{g}(u))J_{\boldsymbol{g}}(u) = D_f(v)^{-1}Q_f(v)^T Q_g(u), D_g(u). \tag{B.37}$$

Since $\boldsymbol{f}(v) = \boldsymbol{g}(u)$ is the same point x on the embedded manifold, the column spaces of $Q_f(v)$ and $Q_g(u)$ form both an orthogonal basis for the tangent space T_x of the embedded manifold at x. In this tangent space they have orthonormal columns (rows) and thus also orthonormal rows (columns). Hence there exists a $d \times d$ orthogonal matrix

$$R(u) := Q_f(v)^T Q_g(u) \tag{B.38}$$

with $R(u)^T R(u) = R(u)R(u)^T = I_d$. Therefore the Jacobian of $\boldsymbol{h}$ admits the pointwise decomposition

$$J_{\boldsymbol{h}}(u) = D_f(v)^{-1} R(u) D_g(u), \tag{B.39}$$

i.e. a diagonal–orthogonal–diagonal factorization valid for every u.

Now, because both p_z and $p_{\hat{z}}$ factorize into independent distributions, introduce the coordinatewise CDF maps (which are smooth, strictly monotone)

$$\tau := (F_{z_1}, ..., F_{z_d}) : \mathbb{R}^d \to (0,1)^d, \qquad \hat{\tau} := (F_{\hat{z}_1}, ..., F_{\hat{z}_d}) : \mathbb{R}^d \to (0,1)^d, \tag{B.40}$$

which push the marginals to independent Unif$(0,1)$ coordinates. Define

$$\psi := \tau \circ \boldsymbol{h} \circ \hat{\tau}^{-1} : (0,1)^d \to (0,1)^d. \tag{B.41}$$

The maps τ and $\hat{\tau}$ are coordinatewise (their Jacobians are diagonal), hence the pointwise structure (B.39) is preserved up to diagonal pre- and post-multiplication: the Jacobian J_ψ has the same diagonal–orthogonal–diagonal form as $J_{\boldsymbol{h}}$ (with possibly different diagonal factors).

Next, we will show that such a mapping ψ will need to be a signed permutation. Recall that $\psi : (0,1)^d \to (0,1)^d$ is a smooth diffeomorphism whose Jacobian admits, at every point, a diagonal–orthogonal–diagonal factorization. This transformation extends continuously and smoothly to a diffeomorphism of the closed hypercube $[0,1]^d$ onto itself. In particular, ψ is a continuous bijection with a continuous inverse on $[0,1]^d$, and thus a homeomorphism. By classical topological results ([17] Domain Invariance Theorem), any homeomorphism between subsets of $\mathbb{R}^d$ maps in-

teriors onto interiors and boundaries onto boundaries. Consequently, ψ maps the hypercube $[0,1]^d$ onto itself and sends boundary points of the hypercube to boundary points. Moreover, since $\boldsymbol{h}$ is globally defined and smooth, the Jacobian field J_ψ satisfies the usual integrability conditions corresponding to a globally defined diffeomorphism. We are therefore exactly in the setting of the Preliminary result in Sec.B.1.1. Applying it to ψ, we conclude that there exists a fixed signed permutation matrix P and smooth, strictly monotone univariate functions $\varphi_i : [0,1] \to [0,1]$ such that

$$\psi(u_1, \ldots, u_d) = P\big(\varphi_1(u_1), \ldots, \varphi_d(u_d)\big), \tag{B.42}$$

for all $(u_1, \ldots, u_d) \in (0,1)^d$. In other words, up to a fixed signed permutation of coordinates, ψ is a global coordinate-wise reparameterization of the unit hypercube.

We now invoke the statistical definition of ψ: by construction, ψ pushes each marginal of the original distribution to the uniform distribution on $(0,1)$. Formally, if $U = \psi(X)$ and $X = (X_1, X_2, X_3)$ denotes a random vector with continuous marginals, then for each $i \in \{1,2,3\}$,

$$U_i = \psi_i(X) \sim \mathrm{Unif}(0,1), \tag{B.43}$$

and the mapping ψ is monotone and invertible along each coordinate.

From the geometric analysis above, we have established that ψ must take the globally diagonal form

$$\psi(u_1, u_2, u_3) = \big(\varphi_1(u_{\sigma(1)}),\ \varphi_2(u_{\sigma(2)}),\ \varphi_3(u_{\sigma(3)})\big), \tag{B.44}$$

for some smooth, strictly increasing univariate functions $\varphi_i : [0,1] \to [0,1]$ and some permutation σ of $\{1,2,3\}$ (encoding the fixed signed permutation of coordinate axes identified earlier).

Because ψ maps each marginal distribution to the uniform law on $(0,1)$, each component φ_i must itself transform a $\mathrm{Unif}(0,1)$ random variable into another $\mathrm{Unif}(0,1)$ random variable. The only smooth, strictly monotone maps on $(0,1)$ satisfying this invariance are the identity and its reflections $t \mapsto 1-t$, corresponding to the two possible signed orientations. Hence,

$$\varphi_i(t) \in \{\, t,\ 1-t \,\}, \qquad \forall i \in \{1,2,3\}. \tag{B.45}$$

Therefore, under the statistical and geometric constraints, the diffeomorphism ψ reduces to a *signed permutation of the coordinates*.

Now, undoing the marginal transforms (i.e. returning from uniform coordinates to the original coordinates) gives

$$\boldsymbol{h} = \tau^{-1} \circ \psi \circ \hat{\tau} = \tau^{-1} \circ P \circ \hat{\tau} \tag{B.46}$$

Hence, $\boldsymbol{h}$ acts as a permutation of coordinates composed with elementwise reparameterizations. In other words, $\boldsymbol{h}$ operates independently on each coordinate of $\hat{\boldsymbol{z}}$ through a unitwise reparameterization $\boldsymbol{t}$ (given by the composition of τ and $\hat{\tau}$, up to permutation) s.t.

$$p_{\boldsymbol{z}} = \boldsymbol{h}_* p_{\hat{\boldsymbol{z}}} = \boldsymbol{t}_* p_{\hat{\boldsymbol{z}}}. \tag{B.47}$$

It is important to note that this is not equivalent to $\boldsymbol{z} = \boldsymbol{t}(\hat{\boldsymbol{z}})$. Indeed, the equality is between distributions. $\boldsymbol{h}$ acts as a unitwise reparameterization on $\hat{\boldsymbol{z}}$ such that the distributions match and thus the axis could still be unaligned. However, because both $p_{\boldsymbol{z}}$ and $p_{\hat{\boldsymbol{z}}}$ are factorized into independent components and because a unitwise reparameterization $\boldsymbol{t}$ preserves independence, the transformed variables $\boldsymbol{t}(\hat{\boldsymbol{z}})$ remain mutually independent. And thus, there is at most an additional global orthogonal transformation Q acting on the coordinates,

$$\boldsymbol{z} = Q \circ \boldsymbol{t}(\hat{\boldsymbol{z}}). \tag{B.48}$$

This ambiguity corresponds to the well-known indeterminacy in Independent Component Analysis (ICA), where the independent components are recoverable only up to scaling, permutation, and orthogonal rotation when multiple Gaussian sources are present.

However, under the additional assumption that at most one latent variable follows a Gaussian distribution, the ICA indeterminacy reduces to permutations and elementwise reparameterizations only. In this case, the orthogonal rotation Q must necessarily be a permutation matrix $\boldsymbol{P}$, since any non-trivial rotation would induce statistical dependence among the coordinates, contradicting the independence of $p_{\boldsymbol{z}}$. Consequently, we obtain

$$\boldsymbol{z} = \boldsymbol{P} \circ \boldsymbol{t}(\hat{\boldsymbol{z}}), \tag{B.49}$$

which is precisely the form of identifiability stated in Def.4.1.

Therefore, the generative process $(\boldsymbol{f}, p_{\boldsymbol{z}})$ is identifiable up to permutation and coordinate-wise invertible reparameterizations of the latent variables, completing the proof. □

Discussion. This proof makes explicit that Prop.4.1 can be viewed as a nonlinear generalization of classical *Independent Component Analysis* (ICA), where identifiability emerges from the interplay between a geometric constraint on the mixing function and a statistical constraint on the latent distribution. In linear ICA, identifiability follows from the fact that a linear mixing matrix with orthogonal columns preserves angles, while statistical independence restricts the admissible transformations to permutations and scalings, except in the presence of multiple Gaussian sources. Our analysis shows that the same logic extends to the nonlinear setting when the Jacobian of the generative mapping satisfies a pointwise orthogonality constraint.

From a geometric perspective, Asm.4.1 enforces a strong form of functional independence: at every point in latent space, each latent coordinate influences the observations along an orthogonal direction in the tangent space of the data manifold. This local orthogonality plays the role of a nonlinear analogue of linear orthogonal mixing. The key technical step of the proof shows that any alternative generative explanation inducing the same observed distribution must differ from the true one by a diffeomorphism whose Jacobian admits a diagonal–orthogonal–diagonal factorization everywhere. Such mappings form a highly restricted class, and when combined with global statistical independence, they collapse to coordinate-wise transformations up to permutation.

The use of marginal CDF transforms highlights this collapse particularly clearly. By reducing the problem to diffeomorphisms of the unit hypercube that preserve independence and admit the same Jacobian structure, the identifiability question becomes purely geometric and topological. The result that such transformations must be signed permutations with coordinate-wise reparameterizations mirrors the classical ICA argument that only trivial symmetries preserve independence. The additional assumption that at most one latent variable is Gaussian then removes the residual orthogonal ambiguity, exactly as in linear ICA.

Importantly, this argument clarifies why nonlinear ICA is generally non-identifiable in the absence of further constraints, as shown in prior impossibility results, and why Asm.4.1 fundamentally changes the situation. Rather than attempting to recover independent components from arbitrary nonlinear mixtures, the orthogonal Jacobian assumption restricts the functional class of admissible generative models to those that behave locally like orthogonal linear maps. In this sense, Asm.4.1 can be interpreted as a structural prior that collapses the nonlinear ICA problem back into an

identifiable regime.

Conceptually, Prop.4.1 demonstrates that identifiability does not rely on linearity per se, but on the preservation of orthogonal functional influence combined with a global statistical constraint. This perspective unifies linear ICA and the more general orthogonal-Jacobian setting studied here under a single principle: disentanglement becomes identifiable whenever local geometric independence and global statistical independence act in concert. This insight also explains why empirical methods that enforce orthogonality or decorrelation in the Jacobian often succeed in practice, even in nonlinear models, when the underlying generative process approximately satisfies these assumptions.

B.1.3 Proof of Proposition 4.2

Proposition 4.2. *Let the observed data $\boldsymbol{x} \in \mathbb{R}^n$ lie on a d-dimensional manifold generated by latent variables $\boldsymbol{z} \in \mathbb{R}^d$, with $d \ll n$, through a smooth and invertible mapping $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^n$ satisfying Asm.4.1. Suppose further that the latent distribution $p_{\boldsymbol{z}}$ satisfies Asm.4.2. Then, the generative process $(\boldsymbol{f}, p_{\boldsymbol{z}})$ is identifiable in the sense of Def.4.1.*

Proof. Assume $\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^n$ is a smooth diffeomorphism satisfying Asm.4.1 (the columns of its Jacobian are orthogonal) and that

$$\boldsymbol{f}_* p_{\boldsymbol{z}} = \boldsymbol{g}_* p_{\hat{\boldsymbol{z}}} \tag{B.50}$$

where $p_{\hat{\boldsymbol{z}}}$ satisfies Asm.4.2. We must show that $\boldsymbol{g}$ recovers $\boldsymbol{f}$ up to the ambiguities allowed in Def.4.1 (a permutation and element-wise reparameterizations).

Because $\boldsymbol{f}$ and $\boldsymbol{g}$ are diffeomorphisms onto the same d-dimensional manifold, define

$$\boldsymbol{h} := \boldsymbol{f}^{-1} \circ \boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^d. \tag{B.51}$$

By the pushforward identity we have

$$\boldsymbol{h}_* p_{\hat{\boldsymbol{z}}} = p_{\boldsymbol{z}}. \tag{B.52}$$

Thus $\boldsymbol{h}$ is a smooth bijection that transforms the density $p_{\hat{\boldsymbol{z}}}$ into the density $p_{\boldsymbol{z}}$.

Fix $u \in \mathbb{R}^d$ and let $v = \boldsymbol{h}(u)$. By Asm.4.1, at the corresponding points

the Jacobians of $\boldsymbol{f}$ and $\boldsymbol{g}$ admit the column-orthogonal factorization

$$J_f(v) = Q_f(v)D_f(v), \qquad J_g(u) = Q_g(u)D_g(u), \tag{B.53}$$

where for each argument $Q_f(v), Q_g(u) \in \mathbb{R}^{n\times d}$ have orthonormal columns ($Q^TQ = I_d$), and $D_f(v), D_g(u) \in \mathbb{R}^{d\times d}$ are diagonal matrices. Because $\boldsymbol{f}$ is a local diffeomorphism, the derivative of $\boldsymbol{f}^{-1}$ at $x = \boldsymbol{f}(v)$ equals the left inverse of $J_f(v)$:

$$D(\boldsymbol{f}^{-1})(x) = (J_f(v))^+ = (J_f(v)^TJ_f(v))^{-1}J_f(v)^T = D_f(v)^{-1}Q_f(v)^T \tag{B.54}$$

where we used $J_f = Q_fD_f$ and $Q_f^TQ_f = I_d$. Consequently, the Jacobian of $\boldsymbol{h}$ at u is

$$J_h(u) = D(\boldsymbol{f}^{-1})(\boldsymbol{g}(u))J_g(u) = D_f(v)^{-1}Q_f(v)^TQ_g(u), D_g(u). \tag{B.55}$$

Since $\boldsymbol{f}(v) = \boldsymbol{g}(u)$ is the same point x on the embedded manifold, the column spaces of $Q_f(v)$ and $Q_g(u)$ form both an orthogonal basis for the tangent space T_x of the embedded manifold at x. In this tangent space they have orthonormal columns (rows) and thus also orthonormal rows (columns). Hence there exists a $d \times d$ orthogonal matrix

$$R(u) \coloneqq Q_f(v)^TQ_g(u) \tag{B.56}$$

with $R(u)^TR(u) = R(u)R(u)^T = I_d$. Therefore the Jacobian of $\boldsymbol{h}$ admits the pointwise decomposition

$$J_h(u) = D_f(v)^{-1}R(u)D_g(u), \tag{B.57}$$

i.e. a diagonal–orthogonal–diagonal factorization valid for every u.

Now, because both p_z and $p_{\hat{z}}$ satisfy Asm.4.2, there exist smooth, strictly monotone coordinate-wise reparameterizations mapping their supports to the unit interval. More precisely, Asm.4.2 ensures the existence of smooth diffeomorphisms

$$\tau \coloneqq (\tau_1, \dots, \tau_d) : \mathbb{R}^d \to (0,1)^d, \qquad \hat{\tau} \coloneqq (\hat{\tau}_1, \dots, \hat{\tau}_d) : \mathbb{R}^d \to (0,1)^d, \tag{B.58}$$

where each $\tau_i, \hat{\tau}_i : \mathbb{R} \to (0,1)$ is strictly monotone. These maps are not assumed to be unique, nor to push p_z and $p_{\hat{z}}$ to the same distribution on $(0,1)^d$; they merely provide a common normalized range.

Define the mapping

$$\psi := \tau \circ \boldsymbol{h} \circ \hat{\tau}^{-1} : (0,1)^d \to (0,1)^d. \tag{B.59}$$

By construction, ψ is a smooth diffeomorphism of the open unit hypercube. Since both τ and $\hat{\tau}$ act independently on each coordinate, their Jacobians are diagonal everywhere. Consequently, conjugation by τ and $\hat{\tau}$ preserves the diagonal–orthogonal–diagonal structure of the Jacobian (with possibly different diagonal factors).

Next, we will show that such a mapping ψ will need to be a signed permutation and coordinate-wise reparametrization. Recall that $\psi : (0,1)^d \to (0,1)^d$ is a smooth diffeomorphism whose Jacobian admits, at every point, a diagonal–orthogonal–diagonal factorization. This transformation extends continuously and smoothly to a diffeomorphism of the closed hypercube $[0,1]^d$ onto itself. In particular, ψ is a continuous bijection with a continuous inverse on $[0,1]^d$, and thus a homeomorphism. By classical topological results ([17] Domain Invariance Theorem), any homeomorphism between subsets of $\mathbb{R}^d$ maps interiors onto interiors and boundaries onto boundaries. Consequently, ψ maps the hypercube $[0,1]^d$ onto itself and sends boundary points of the hypercube to boundary points. Moreover, since $\boldsymbol{h}$ is globally defined and smooth, the Jacobian field J_ψ satisfies the usual integrability conditions corresponding to a globally defined diffeomorphism. We are therefore exactly in the setting of the Preliminary result in Sec.B.1.1. Applying it to ψ, we conclude that there exists a fixed signed permutation matrix P and smooth, strictly monotone univariate functions $\varphi_i : [0,1] \to [0,1]$ such that

$$\psi(u_1, \ldots, u_d) = P\big(\varphi_1(u_1), \ldots, \varphi_d(u_d)\big), \tag{B.60}$$

for all $(u_1, \ldots, u_d) \in (0,1)^d$. In other words, *up to a fixed signed permutation of coordinates, ψ is a global coordinate-wise reparameterization of the unit hypercube.*

Now, undoing the marginal normalizations (i.e. returning to the original coordinates) yields

$$\boldsymbol{h} = \tau^{-1} \circ \psi \circ \hat{\tau} = \tau^{-1} \circ P \circ \phi \circ \hat{\tau}, \tag{B.61}$$

where P is a fixed permutation matrix and $\phi(u_1, \ldots, u_d) = (\phi_1(u_1), \ldots, \phi_d(u_d))$ is a coordinate-wise smooth, strictly monotone reparameterization of $(0,1)^d$.

Consequently, $\boldsymbol{h}$ acts as a permutation of coordinates composed with element-wise reparameterizations. Equivalently, there exists a coordinate-wise invertible map $\boldsymbol{t} : \mathbb{R}^d \to \mathbb{R}^d$ such that

$$p_{\boldsymbol{z}} = \boldsymbol{h}_* p_{\hat{\boldsymbol{z}}} = \boldsymbol{t}_* p_{\hat{\boldsymbol{z}}}. \tag{B.62}$$

It is important to emphasize that this equality holds at the level of distributions: it does not imply the pointwise identity $\boldsymbol{z} = \boldsymbol{t}(\hat{\boldsymbol{z}})$. Indeed, a unit-wise reparameterization may map distributions correctly even if the coordinate axes of $\boldsymbol{z}$ and $\hat{\boldsymbol{z}}$ are not aligned.

However, both $p_{\boldsymbol{z}}$ and $p_{\hat{\boldsymbol{z}}}$ satisfy Asm.4.2. In particular, since they have bounded support that are hypercubes up to a reparametrization, their supports must coincide with a compact hyperrectangle, and their copulas have full support on the interior of this domain. As a consequence, the diffeomorphism $\boldsymbol{h}$ extends continuously to the boundary of the support and maps boundary points to boundary points and interior points to interior points. Under these conditions, any smooth bijection mapping $p_{\hat{\boldsymbol{z}}}$ to $p_{\boldsymbol{z}}$ through a coordinate-wise reparameterization must preserve the axis-aligned structure of the support. Indeed, any nontrivial mixing of coordinates would necessarily map a boundary face of the hyperrectangle to a set intersecting the interior, or would destroy the full-support property of the copula, contradicting Asm.4.2. Therefore, the only admissible transformations are coordinate-wise reparameterizations composed with permutations of the coordinate axes.

We can thus conclude that $\boldsymbol{h}$ is identifiable up to a permutation and coordinate-wise invertible reparameterizations of the latent variables. Consequently, the generative process $(\boldsymbol{f}, p_{\boldsymbol{z}})$ is identifiable in the sense of Def.4.1, completing the proof. □

Discussion. The proof of Prop.4.2 makes explicit the mechanisms underlying identifiability in the dependent case, and in particular clarifies the respective roles played by support geometry and distributional assumptions. Building on these insights, the full-support copula assumption can be viewed as a natural generalization of the independence setting: requiring that all combinations of latent coordinates occur ensures that no factor is functionally constrained by the others, which is precisely what enables disentanglement. However, in contrast to the independent case, this assumption alone is not sufficient unless the latent variables are supported on a *bounded* domain.

This additional requirement is essential for identifiability. When the latent space is $\mathbb{R}^d$, there exist infinitely many smooth, monotone transformations that map $\mathbb{R}^d$ onto a bounded domain such as the hypercube $(0,1)^d$. For instance, applying sigmoid-type transformations along arbitrary directions yields different embeddings of an unbounded domain into the same bounded support, all of which are compatible with the same observed distribution. As a result, the geometry of the latent space cannot be uniquely recovered from data in this setting.

By contrast, when the latent variables are supported on a bounded domain with full copula support, the boundary structure of the support becomes identifiable and constrains the class of admissible transformations, leading to disentanglement up to coordinate-wise reparametrization and permutation. The independent case does not require such a boundedness assumption because independence provides strictly stronger information: it fixes not only the support geometry but also the factorization structure of the distribution, thereby ruling out these additional ambiguities even when the latent space is $\mathbb{R}^d$.

B.1.4 Discussion on Assumption 4.1 and Assumption 4.1

This section elaborates on the practical scope and interpretation of Asm.4.1 and 4.2. These assumptions are not only theoretically motivated but also well-grounded in widely accepted principles and common practices in representation learning; we therefore view them as structural abstractions that enable studying identifiability beyond the classical independence setting. In what follows, we discuss each assumption in turn, clarifying its conceptual grounding, its relationship to causal representation learning settings, and the conditions under which it can be expected to hold in practice.

Assumption 4.1 (Orthogonal Jacobian). Asm.4.1 should be understood not as an empirical property to be verified in a given dataset, but as a *definitional criterion* for what constitutes a meaningful latent concept: each factor contributes a functionally independent mode of variation, decoupled from all others in its local effect on the observations. This view is strongly aligned with the Independent Causal Mechanisms (ICM) principle from causal inference [134], a widely accepted foundation which posits that the mechanisms governing the components of a system operate autonomously and do not inform one another. Applied outside a strictly causal context, Asm.4.1 can be understood as a functional, non-parametric instantiation

of ICM: instead of requiring independence in a causal graph, it requires decoupling in the generative Jacobian. This connection is well-supported by the theoretical analysis in [52], where orthogonal Jacobian structure was introduced precisely to capture this notion.

Beyond its theoretical motivation, Asm.4.1 finds direct empirical support in practical settings. Wei *et al.* [159] successfully leveraged Jacobian orthogonality constraints for unsupervised disentanglement in image generation, demonstrating that such structural priors are not only theoretically meaningful but also empirically effective on realistic data. Moreover, as discussed in Sec.4.6, VAEs implicitly encourage a related structure through their factorized approximate posterior [129, 5], which helps explain their empirical disentanglement ability on structured datasets.

In summary, Asm.4.1 is best understood not as an empirical property to be verified in any given dataset, but as a *structural design criterion* for what kind of representation one seeks to learn: one in which each factor's functional influence on the observations is orthogonal to that of all others.

Assumption 4.2 (Full-Support Copula). Asm.4.2 formalizes the combinatorial richness of the latent factors, capturing the idea that each factor should, in principle, be independently manipulable. Many latent variables (e.g., position, orientation, lighting) naturally vary across bounded ranges, and datasets are often explicitly constructed to explore these variations, making this assumption well-aligned with common practices in representation learning.

A fundamental distinction, however, is that Asm.4.2 is a *dataset-level property* of the support of the latent distribution, rather than a property of any underlying causal or generative model. The existence of a causal model does not, by itself, guarantee that the resulting dataset will be rich enough to satisfy it. To illustrate this, consider two latent factors (Z_1, Z_2) with a causal dependency $Z_1 \to Z_2$:

- *Deterministic mechanism*: if $Z_1 = 0$ implies $Z_2 = 0$ and $Z_1 = 1$ implies $Z_2 = 1$, then combinations such as $(Z_1 = 0, Z_2 = 1)$ never occur. The dataset lacks full combinatorial coverage, violating Asm.4.2, even though a valid causal model is present.
- *Probabilistic mechanism*: if $Z_1 = 0$ yields $P(Z_2 = 0) = 0.9$ and $P(Z_2 = 1) = 0.1$, then every combination of (Z_1, Z_2) appears with non-zero probability. Causal dependencies exist and the variables are statistically dependent, yet Asm.4.2 is perfectly satisfied because the joint latent space has full support.

This shows that Asm.4.2 can be satisfied in typical causal representation learning settings, provided the dataset is sufficiently diverse. Moreover, even when natural observational data lacks certain combinations, interventions (either explicit or through deliberate dataset design) can restore the required combinatorial richness.

Crucially, this highlights that causal structure and Asm.4.2 are *orthogonal*: our work does not rely on any causal assumptions, but rather focuses on the combinatorial richness of the dataset, regardless of how the data was generated (causally or not). This formalizes the idea that each latent factor should correspond to a distinct, manipulable degree of freedom ("concept slider"). If two variables can never be varied separately in the observed data, they cannot be identified as separate entities without further supervision. Thus, while Asm.4.2 defines the conditions under which disentanglement is theoretically possible, it may not be achieved in every real-world scenario, particularly if the dataset is not sufficiently rich or lacks the necessary interventional samples to make these separate degrees of freedom visible to the disentanglement model. For instance, variables such as temperature and altitude may not span all possible combinations in observed data, even though such combinations are physically plausible. So the assumption should be interpreted as a property of the underlying generative domain, not a guarantee about any finite dataset. Even so, the assumption provides a critical theoretical foundation: it delineates when disentanglement is possible, by ensuring that distinct latent factors are identifiable in principle. In other words, while actual datasets may be incomplete, Asm.4.2 highlights the necessary richness required to make independent degrees of freedom visible to the model, grounding identifiability results in realistic, physically plausible factor spaces.

B.2 Disentanglement Under General Latent Domains

In the main text, we argued that identifiability does not necessarily require strong distributional assumptions such as statistical independence. Instead, *limited but appropriate global information about the latent domain* may already suffice to fully determine the generative mapping. In particular, when the latent variables are supported on a known,bounded, simply connected domain $\Omega \subset \mathbb{R}^d$, the generative process can become identifiable up to the natural rigid symmetries of that support. In the main body of the pa-

per, this idea was instantiated through Asm.4.2, which restricts the latent domain (up to coordinate-wise reparameterization) to a hypercube. However, the above observation suggests that identifiability can be achieved under substantially more general geometric conditions. The purpose of this appendix is to formalize and extend this claim by studying disentanglement when the latent domain is known but not necessarily a hypercube.

For clarity, we restrict our analysis in this appendix to the class of *conformal* generative mappings. This setting allows us to leverage classical results from geometric function theory and to make precise statements about how domain geometry constrains admissible transformations. In particular, for conformal mappings in dimensions $d \geq 3$, the knowledge of a small amount of global geometric information—such as the support of the latent variables or the correspondence of a few salient points—is already sufficient to uniquely determine the mapping up to rigid symmetries.

This intuition is formalized in the following proposition, which establishes identifiability under mild geometric assumptions on the latent support.

Proposition B.1. *Let the observed data $\boldsymbol{x} \in \mathbb{R}^n$ lie on a d-dimensional manifold generated by latent variables $\boldsymbol{z} \in \mathbb{R}^d$, with $3 \leq d \ll n$, through a smooth and invertible mapping $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^n$ that is conformal everywhere on the domain in an open subset. Suppose further that the latent variables follow a distribution $p_{\boldsymbol{z}}$ supported on a known, bounded, simply connected region $\Omega \subset \mathbb{R}^d$. For any $c \in \mathbb{R}^d$, there exist $u, v \in \mathbb{S}^{d-1}$ such that, denoting by $(r_1(u), r_2(u))$ and $(r_1(v), r_2(v))$ the intersections of Ω with the rays emanating from c, one has $r_1(u)r_2(u) \neq r_1(v)r_2(v)$. Then, the generative process $(\boldsymbol{f}, p_{\boldsymbol{z}})$ is identifiable in the sense of Def.4.1, up to a rigid transformation that is a symmetry of Ω.*

The key theoretical ingredient underlying Prop.B.1 is Liouville's theorem for conformal mappings in dimensions $d > 2$ [13]. This result implies that any conformal transformation of $\mathbb{R}^d$ must be a Möbius transformation, i.e., a composition of translations, rotations, dilations, and inversions. Among these components, inversion is the only source of nonlinearity. When the latent domain Ω is bounded, simply connected, and not invariant under inversion (a property satisfied by a broad class of supports including all polytopes and most bounded smooth domains), nonrigid Möbius transformations cannot preserve Ω. The geometric condition imposed in Prop.B.1 precisely excludes this possibility by ensuring that no inversion can map Ω onto itself. As a result, the only admissible transfor-

mations are rigid motions (rotations and translations) that are symmetries of the domain.

This establishes that identifiability can be achieved under substantially weaker assumptions than those imposed in the main text. In particular, it shows that disentanglement is possible for a wide class of known latent supports, not only hypercubes, provided that the domain geometry sufficiently constrains admissible conformal symmetries. Only highly symmetric sets, such as perfect spheres, violate this condition and admit non-rigid self-transformations.

We empirically illustrate this phenomenon in Fig.4.4 of the main chapter and in Figs.B.11 and B.12 of App.B.4, where we observe identifiability up to rigid rotations for several non-cubic latent domains.

In the following sections, we provide a proof of Prop.B.1 along with additional geometric intuition and visualizations. We then treat separately the special case $d = 2$, where conformal mappings exhibit fundamentally different behavior.

B.2.1 Proof and Geometric Interpretation

We now prove Prop.B.1 and provide geometric intuition for the role played by the latent domain geometry.

Let $\boldsymbol{f}, \boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^n$ be smooth, invertible mappings that are conformal on an open neighborhood of a bounded, simply connected domain $\Omega \subset \mathbb{R}^d$, with $d \geq 3$. Assume that the induced data distributions coincide, i.e.,

$$\boldsymbol{f}_* p_{\boldsymbol{z}} = \boldsymbol{g}_* p_{\hat{\boldsymbol{z}}}, \tag{B.63}$$

where both $p_{\boldsymbol{z}}$ and $p_{\hat{\boldsymbol{z}}}$ are supported on Ω.

Since $\boldsymbol{f}$ and $\boldsymbol{g}$ are diffeomorphisms onto the same d-dimensional data manifold, we define

$$\boldsymbol{h} := \boldsymbol{f}^{-1} \circ \boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^d. \tag{B.64}$$

By construction, $\boldsymbol{h}$ is a smooth bijection. Using the pushforward identity, we obtain

$$\boldsymbol{h}_* p_{\hat{\boldsymbol{z}}} = p_{\boldsymbol{z}}, \tag{B.65}$$

which implies that $\boldsymbol{h}$ maps Ω onto itself almost everywhere. Hence, $\boldsymbol{h}$ is an automorphism of the latent domain Ω.

Because conformality is preserved under inversion and composition, and both $\boldsymbol{f}$ and $\boldsymbol{g}$ are conformal, the map $\boldsymbol{h}$ is itself conformal on an open

neighborhood of Ω. The identifiability problem therefore reduces to characterizing all conformal automorphisms of Ω.

For dimensions $d \geq 3$, Liouville's theorem [13] states that any conformal map between open subsets of $\mathbb{R}^d$ must be a Möbius transformation. Consequently, $\boldsymbol{h}$ must be of the form

$$\boldsymbol{h}(x) = b + \frac{\alpha A(x-a)}{\|x-a\|^{\varepsilon}}, \tag{B.66}$$

where $a, b \in \mathbb{R}^d$, $\alpha \in \mathbb{R} \setminus \{0\}$, $A \in O(d)$ is an orthogonal matrix, and $\varepsilon \in \{0, 2\}$. Here, a denotes the center of inversion, b a translation, α a uniform dilation, and A a rotation or reflection. The case $\varepsilon = 0$ corresponds to affine conformal maps (no inversion), while $\varepsilon = 2$ corresponds to Möbius transformations involving inversion.

Geometrically, Möbius transformations admit a classical interpretation via stereographic projection. Specifically, they can be obtained by projecting $\mathbb{R}^d$ onto the d-sphere, applying a rigid motion of the sphere, and projecting back via stereographic projection. A visualization of this construction is provided in Fig.B.1. In this representation, inversion corresponds to a transformation that exchanges a point a with infinity: when $\varepsilon = 2$, the point $x = a$ is mapped to infinity, and infinity is mapped to b. When $\varepsilon = 0$, infinity is preserved.

We now analyze which Möbius transformations can be automorphisms of the bounded domain Ω.

First, consider the case $\varepsilon = 0$. In this case, $\boldsymbol{h}$ reduces to an affine conformal map,

$$\boldsymbol{h}(x) = b + \alpha A(x-a). \tag{B.67}$$

Translations, rotations, and dilations preserve the global shape of a domain. Since Ω is bounded, $\boldsymbol{h}$ cannot involve arbitrary translations or dilations while remaining an automorphism. Consequently, $\boldsymbol{h}$ must reduce to a rigid transformation belonging to the symmetry group of Ω. This yields identifiability up to rigid symmetries.

We now turn to the only source of nonlinearity: the case $\varepsilon = 2$, corresponding to inversion. All other components of a Möbius transformation (translations, rotations, and dilations) preserve the qualitative shape of a domain. Therefore, if a Möbius transformation is to act as a nontrivial automorphism of Ω, the inversion component must itself preserve the domain.

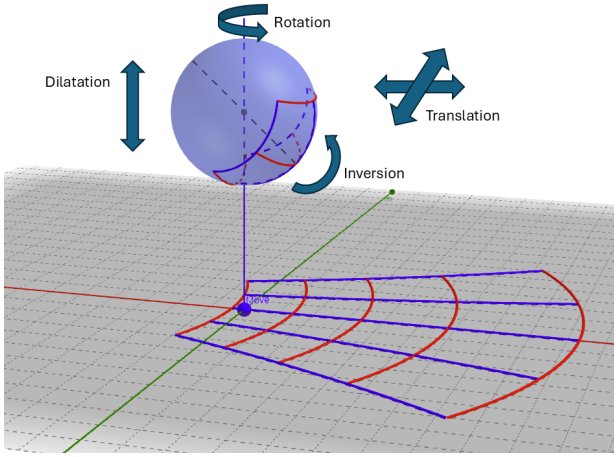


Fig. B.1 Geometric interpretation of Möbius transformations via stereographic projection. Points in $\mathbb{R}^d$ are projected onto the sphere, transformed by a rigid motion of the sphere, and mapped back to $\mathbb{R}^d$ via inverse stereographic projection. Plot generated using a visualization tool created by Juan Carlos Ponce Campuzano on Geogebra

Without loss of generality, we may reduce the analysis to inversion with respect to the unit sphere,

$$\mathcal{I}(x) = \frac{x}{\|x\|^2}, \tag{B.68}$$

since any other inversion can be obtained by composing $\mathcal{I}$ with translations, rotations, and dilations.

Inversion has two fundamental geometric effects. First, it exchanges the interior and exterior of the unit sphere: points inside the sphere are mapped outside, and vice versa. Second, the center of inversion is sent to infinity, and infinity is sent to the center. Since Ω is bounded, the center of inversion must lie outside Ω; otherwise, points in Ω would be mapped to infinity, which is incompatible with $\boldsymbol{h}$ being an automorphism.

For inversion to preserve Ω, the domain must therefore intersect the inversion sphere. Moreover, the portion of Ω inside the sphere must be exactly the inversion image of the portion outside the sphere, while the intersection of Ω with the sphere itself remains invariant. This implies that Ω must cross the inversion sphere in a highly structured way: one part of the

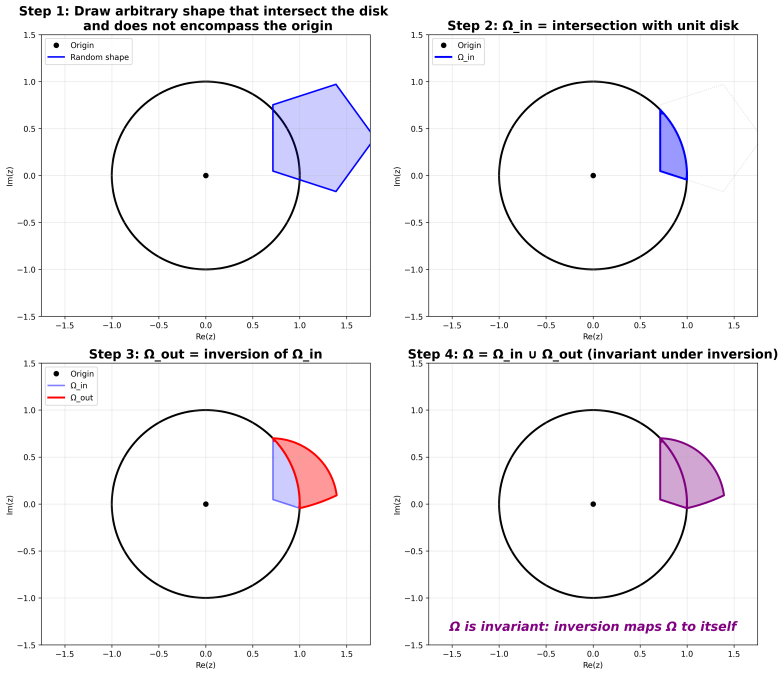


Fig. B.2 Method for constructing counterexamples. Any domain exhibiting inversion symmetry can be generated in this way.

domain lies inside, one part lies outside, and these two parts are mapped onto each other by the inversion. This construction characterizes all possible counterexamples in which a nontrivial inversion acts as a domain automorphism. We visualize such configurations in Fig.B.2, where the interior and exterior components are exact inversion images of one another.

Crucially, aside from the points lying exactly on the inversion sphere, no point can remain fixed. Thus, for any other point in Ω, the existence of an automorphism involving inversion requires the existence of a corresponding point on the opposite side of the inversion sphere. It is precisely this requirement that is excluded by the geometric condition of Prop.B.1.

Indeed, the condition in Prop.B.1 ensures that for any potential inversion center c, there exist directions along which the radial structure of Ω is incompatible with inversion symmetry. Prop.B.1 excludes all domains that are invariant under an inversion. As a result, no inversion can map Ω onto itself. This situation is illustrated in Fig.B.3.

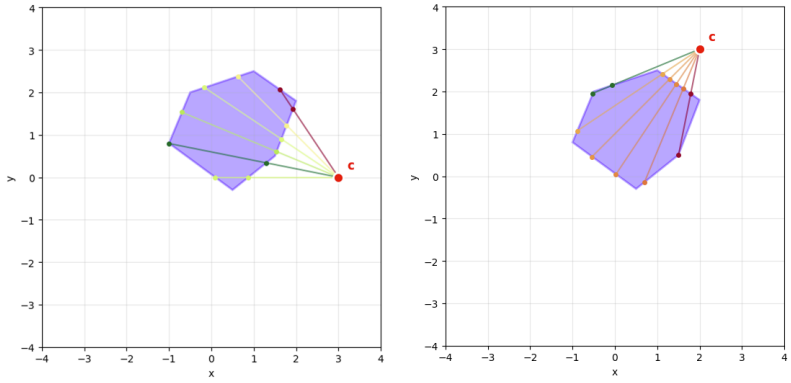


Fig. B.3 The condition in Prop.B.1 directly targets the defining geometric invariant of inversion. For a fixed center c, inversion about a sphere centered at c preserves, along every ray direction u, the product of the two intersection distances $r_1(u)r_2(u)$ with the domain boundary. Domains that are invariant under inversion, such as disks or spheres, are characterized by the fact that this product is constant across all directions. Prop.B.1 explicitly excludes this possibility by requiring the existence of at least two directions for which the radial products differ, thereby breaking the invariance required for inversion symmetry. Since any Möbius transformation involving inversion must preserve this radial structure for some center c, the condition rules out all non-rigid Möbius automorphisms. Consequently, only affine conformal maps can act as automorphisms of the latent domain.

We conclude that the inversion case $\varepsilon = 2$ is ruled out under the assumptions of Prop.B.1. Therefore, any conformal automorphism of Ω must be affine, and hence a rigid symmetry of the domain. It follows that

$$g = f \circ h, \tag{B.69}$$

where h is a rigid transformation preserving Ω. This establishes identifiability of the generative process up to the natural rigid symmetries of the latent domain, completing the proof of Prop.B.1.

Most latent domains do not admit invariance under inversion; only highly specific and strongly symmetric sets, such as perfect spheres or disks, possess this property. For such exceptional geometries, non-rigid Möbius symmetries may persist, leading to multiple valid generative solutions. We illustrate this phenomenon in Fig.B.4 with the case of a disk, whose inversion symmetry yields multiple admissible solutions, and in Fig.B.5 with a crescent-shaped domain, where inversion symmetry gives

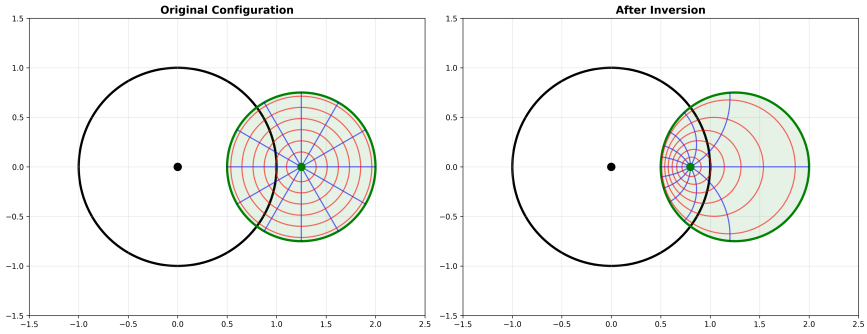


Fig. B.4 Visualization of a disk automorphism induced by an inversion. The disk is mapped onto itself, confirming that the transformation is indeed an automorphism. The mapping is nonlinear and produces a visibly deformed geometry.

rise to exactly two distinct solutions. Importantly, such cases are rare: for the vast majority of bounded, simply connected domains, inversion invariance does not hold. Consequently, identifiability extends well beyond the hypercube setting considered in the main text, demonstrating that disentanglement can be achieved for a broad class of latent supports whenever the domain geometry sufficiently constrains admissible conformal symmetries.

B.2.2 The Two-Dimensional Case

The two-dimensional case exhibits fundamentally different behavior from the higher-dimensional setting discussed above. This distinction arises from the markedly greater flexibility of conformal mappings in dimension $d = 2$, which contrasts sharply with the rigidity imposed by Liouville's theorem for $d \geq 3$.

In two dimensions, conformal mappings coincide with holomorphic or anti-holomorphic functions with non-vanishing derivative. As a result, the space of conformal self-maps of planar domains is infinite-dimensional, rather than being restricted to Möbius transformations as in higher dimensions. This increased flexibility is formalized by the classical *Riemann Mapping Theorem* (RMT) [153].

Riemann Mapping Theorem. The Riemann Mapping Theorem states that any simply connected, proper open subset $\Omega \subset \mathbb{C}$ (i.e., $\Omega \neq \mathbb{C}$) is conformally equivalent to the open unit disk $\mathbb{D}$. That is, there exists a bijective

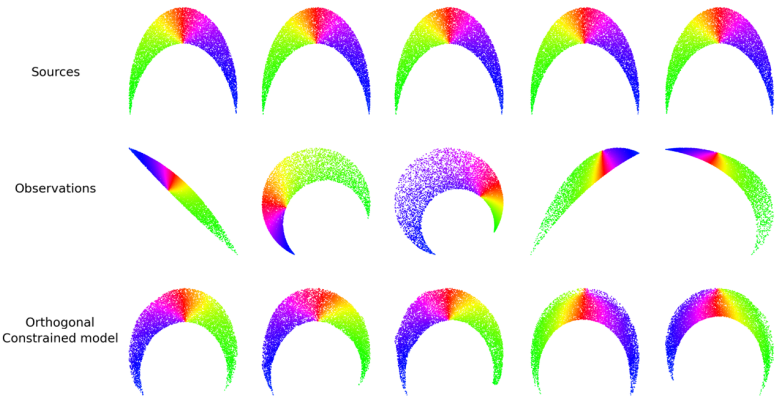


Fig. B.5 Qualitative disentanglement results for crescent-shaped domains. From top to bottom: ground-truth sources sampled from the structured domain, corresponding observations, and reconstructions obtained using the orthogonally constrained model. Due to inversion symmetry, two equivalent solutions are observed.

conformal map

$$\phi : \Omega \to \mathbb{D}. \tag{B.70}$$

Moreover, such a map is unique up to composition with a conformal automorphism of the disk, i.e., a Möbius transformation preserving $\mathbb{D}$.

This result implies that, in dimension two, any simply connected domain can be conformally deformed into a disk, regardless of its geometric shape. Consequently, unlike the case $d \geq 3$, domain geometry alone is insufficient to restrict conformal symmetries to rigid motions. The group of conformal automorphisms of the disk is infinite-dimensional, allowing for highly non-rigid transformations.

At first glance, this suggests that identifiability under conformal generative models should fail in two dimensions. However, this conclusion is overly pessimistic in the present setting, for several important reasons.

1) Role of boundary behavior. The Riemann Mapping Theorem applies to *open* domains and makes no guarantees about the behavior of the conformal map on the boundary $\partial\Omega$. In general, the conformal equivalence provided by the RMT does *not* extend continuously, let alone smoothly, to the boundary. As a result, many conformal maps that exist in the interior of a domain are incompatible with prescribed boundary correspondences.

In our setting, the generative mapping f is assumed to be smooth and invertible on an open neighborhood of the latent domain, which implicitly constrains its behavior near and on the boundary. When the boundary geometry of Ω is taken into account (particularly when it contains salient geometric features such as corners, edges, or extremal points) the class of admissible conformal automorphisms is drastically reduced.

2) Fixing boundary points and rigidity. A classical result in complex analysis states that a conformal automorphism of the unit disk is uniquely determined by the images of three distinct boundary points. Equivalently, any conformal automorphism of a planar domain that fixes three non-collinear points must be the identity.

This observation has direct implications for identifiability. If the latent domain Ω possesses three or more geometrically distinguished boundary points, such as the vertices of a triangle, the corners of a square or rectangle, or other extremal points, then any conformal automorphism preserving these points is necessarily rigid. In such cases, the freedom predicted by the Riemann Mapping Theorem disappears once boundary correspondence is enforced.

For example, if Ω is a polygonal domain, the interior angles at its vertices are conformal invariants.Any conformal automorphism that preserves the domain must therefore map vertices to vertices while preserving angles, which leaves only rigid motions as admissible transformations. Similar reasoning applies to domains with sufficiently rich boundary structure.

Implications for identifiability in practice. Although two-dimensional conformal geometry admits greater theoretical flexibility, the combination of smoothness, invertibility, and implicit boundary constraints imposed by the generative model typically rules out non-rigid conformal automorphisms. As a result, even in $d = 2$, identifiability often holds up to rigid transformations for most practically relevant latent domains.

In this sense, the two-dimensional case represents a boundary regime: while conformal maps are theoretically more expressive, the conditions required for non-identifiability are fragile and rarely satisfied in structured or bounded domains. Thus, once boundary regularity or correspondence is taken into account, the identifiability conclusions obtained for $d \geq 3$ extend, in practice, to the two-dimensional setting as well.

This completes the discussion of the $d = 2$ case and highlights that the apparent loss of rigidity predicted by the Riemann Mapping Theorem does not fundamentally undermine identifiability in the present framework.

B.3 Experimental Setup

This appendix details the experimental setup used throughout Sec.4.5, including dataset generation, ground-truth mixing functions, model architectures, training objectives, and evaluation metrics. All experiments are implemented in Python using PyTorch [119], leveraging automatic differentiation to compute full Jacobians and enforce orthogonality constraints during training.

B.3.1 Datasets and Generative Processes

To make the dataset for the experiments we start by generating data sources. We consider latent variables $\boldsymbol{z} \in \mathbb{R}^d$ with $d \in \{3, 6, 9\}$ and generate datasets under two distinct regimes: independent and dependent factors.

In the **independent setting**, latent variables are sampled i.i.d. from a uniform distribution on the hypercube,

$$\boldsymbol{z} \sim \mathcal{U}([0,1]^d). \tag{B.71}$$

In the **dependent setting**, we generate complex statistical dependencies while preserving Asm.4.2 (full-support copula). We first sample $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ and transform it using an untrained autoregressive normalizing flow composed of 20 RealNVP-style coupling layers [36]. A coordinate-wise sigmoid nonlinearity maps the samples to $[0,1]^d$. To ensure full support, we mix this distribution with a uniform component:

$$p_{\boldsymbol{z}} = \alpha\, p_{\mathrm{NF}} + (1-\alpha)\, \mathcal{U}([0,1]^d), \tag{B.72}$$

with $\alpha \in [0,1]$. This guarantees a minimum density of $1-\alpha$ everywhere on the hypercube, hence full support. Typically $\alpha = 0.75$ is chosen. This construction explicitly violates independence while satisfying Asm.4.2. For each configuration, we sample 10 000 latent points.

Once we have the sources, latent variables are then mapped to observations $\boldsymbol{x} \in \mathbb{R}^n$ using smooth, invertible transformations satisfying Asm.4.1.

We consider two classes of mixing functions. First, we use **conformal mappings** based on generalized Möbius transformations,

$$f(\boldsymbol{z}) = \boldsymbol{b} + \frac{\alpha\, A(\boldsymbol{z} - \boldsymbol{a})}{\|\boldsymbol{z} - \boldsymbol{a}\|^{\varepsilon}}, \tag{B.73}$$

where A is orthogonal, $\alpha > 0$, $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$, and $\varepsilon = 2$. This choice yields non-isometric but conformal transformations, preserving angles while allowing local scaling.

Second, to test identifiability beyond the conformal case, we construct a broad class of nonlinear mixing functions whose Jacobians are of the form $J_f(\boldsymbol{z}) = Q(\boldsymbol{z})D(\boldsymbol{z})$, but which are not globally conformal. These mappings are generated compositionally, exploiting the fact that the composition of QD transformations remains QD if composed block-wise.

Concretely, we partition the latent space into blocks of three dimensions. Within each block, we first apply a nonlinear two-dimensional transformation acting on the first two coordinates, chosen among polar, elliptic, parabolic, or rigid transformations. This 2D mapping is then composed with a 2D Möbius transformation, yielding a flexible yet analytically tractable deformation with orthogonal Jacobian structure. The resulting 2D surface is subsequently embedded into three dimensions, either by directly appending the third coordinate or by constructing a surface of revolution around it. In the latter case, a point (u, v, ϕ) is mapped according to

$$x = X(u, v)\cos\phi, \qquad y = X(u, v)\sin\phi, \qquad z = Z(u, v), \tag{B.74}$$

where $(X(u, v), Z(u, v))$ denotes the output of the preceding 2D transformation. Each three-dimensional block is then further composed with a 3D Möbius transformation. After processing all blocks independently, we apply an additional global Möbius transformation across all dimensions to induce cross-block interactions and ensure full mixing. At each stage, transformation parameters are sampled within bounded ranges to avoid singularities (e.g., degeneracies associated with Möbius inversions), and intermediate outputs are normalized to ensure comparable scales across dimensions. Examples of transformations from the hypercube are provided in Fig.B.6

This construction yields highly expressive nonlinear mixing functions with orthogonal Jacobians that go far beyond simple coordinate reparameterizations such as spherical or cylindrical coordinates. While Möbius transformations alone form a relatively restricted subclass, their block-wise and hierarchical composition produces a much larger family of admissible QD mappings. As a result, these generative processes are substantially more challenging to disentangle than purely conformal ones, providing a stringent empirical test of the proposed identifiability theory.

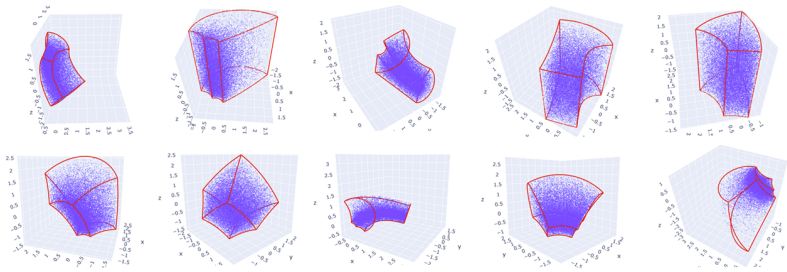


Fig. B.6 Example of transformations in 3D.

B.3.2 Evaluation Metrics

We evaluate both density estimation accuracy and disentanglement quality.

Since the true data-generating distribution is known, we compute the **Kullback–Leibler divergence** between the true distribution and the learned model by evaluating likelihoods using the inverse ground-truth mapping. Concretely, since we have access to the ground-truth mixing function f and can compute its Jacobian via automatic differentiation, the true data log-density is tractable at any point. The learned normalizing flow provides a tractable model density. The KL divergence is then estimated by a standard Monte Carlo average over samples from the true distribution. This metric purely assesses goodness of fit and is not used to measure disentanglement. All models achieve comparable KL values, confirming successful density estimation.

Disentanglement is assessed using the **Mean Correlation Coefficient (MCC)** [83]. We compute the pairwise correlation matrix between ground-truth latent variables and recovered representations, solve a maximum-weight assignment problem, and report the mean of the matched correlations. We use Spearman correlation [57], which is invariant to monotonic coordinate-wise reparameterizations, consistent with Def.4.1.

We also report the **nonlinear Amari distance** introduced in [52] that directly compares the Jacobians of the true mixing and the learned unmixing. Given the true mixing f and learned unmixing g, it is defined as

$$d_{\text{n-Amari}}(g, f) = \mathbb{E}_{\boldsymbol{x}\sim p_x}\Big[d_{\text{Amari}}\big(J_g(\boldsymbol{x})\, J_f(f^{-1}(\boldsymbol{x}))\big)\Big]. \tag{B.75}$$

where d_{Amari} is the classical Amari distance metric used in linear ICA [6]. This metric equals zero if and only if $g \circ f$ differs from the identity by a permutation and coordinate-wise reparameterization (i.e., the product of Jacobians is a scale and permutation matrix), exactly matching the equivalence class of Def.4.1. Together, MCC and nonlinear Amari distance quantify whether the learned representation recovers the true factors up to the admissible ambiguities.

B.3.3 Models, Architectures, and Training Objectives

Given observations $\boldsymbol{x}$, the objective is to recover latent representations corresponding to the unknown sources, without access to the true generative factors. To this end, we train expressive normalizing flows on the observed data and evaluate whether the resulting latent variables exhibit a disentangled structure.

Our main architecture is a **Residual Flow** [26], defined by the transformation

$$\boldsymbol{y} = f(\boldsymbol{x}) = \boldsymbol{x} + h(\boldsymbol{x}), \tag{B.76}$$

where $h : \mathbb{R}^d \to \mathbb{R}^d$ is a neural network constrained to be Lipschitz-continuous, which guarantees invertibility and ensures that the Jacobian

$$J_f(\boldsymbol{x}) = I + J_h(\boldsymbol{x}) \tag{B.77}$$

is full and expressive. Residual flows are universal density approximators and allow for highly flexible transformations while retaining tractable log-determinants.

We use 64 residual flow layers, each parameterized by a 3-layer multilayer perceptron with hidden dimension 128. Models are trained using Adam with learning rate 10^{-3} and batch size 256 until convergence.

We consider two settings depending on whether the latent sources are independent or dependent. In the *independent* setting, the base distribution of the residual flow is a logistic distribution with diagonal scale matrix,

$$p_{\text{base}}(\boldsymbol{z}) = \text{Logistic}(\boldsymbol{0}, I), \tag{B.78}$$

which provides full support on $\mathbb{R}^d$ and matches standard assumptions in nonlinear ICA.

In the *dependent* setting, we allow the base distribution to be complex and learned. Specifically, we train an additional normalizing flow to model the prior distribution. This prior flow consists of three affine coupling layers, each implemented by a 3-layer MLP with hidden size 64. To guarantee full support and avoid degeneracies, we define the base distribution as a mixture

$$p_{\text{base}} = \beta\, p_{\text{logistic}} + (1-\beta)\, p_{\text{prior NF}}, \qquad \beta \in (0,1). \tag{B.79}$$

This construction exploits the fact that pushforwards preserve mixtures:

$$f_{\#}(\beta P + (1-\beta)Q) = \beta\, f_{\#}P + (1-\beta)\, f_{\#}Q. \tag{B.80}$$

As a result, the learned latent distribution is a mixture between a flexible learned component and a simple reference distribution. Intuitively, after applying a sigmoid map, the logistic component corresponds to a uniform distribution on $[0,1]^d$, ensuring non-vanishing density everywhere. This guarantees full support in the latent space and therefore enforces Asm.4.2 inherently during training.

It is important to note that the base distribution lives on $\mathbb{R}^d$, whereas the true sources are supported on $[0,1]^d$. Consequently, even in the ideal case, recovery is only possible up to coordinate-wise reparameterizations. The learned latent variables are therefore not expected to match the source distribution directly. Sigmoid mappings to $[0,1]^d$ are applied solely for visualization purposes and are never used during training.

Unconstrained models are trained by maximizing the standard log-likelihood

$$\log p(\boldsymbol{x}) = \log p(g(\boldsymbol{x})) + \log|\det J_g(\boldsymbol{x})|. \tag{B.81}$$

Orthogonally constrained models use the regularized objective proposed by Gresele *et al.* [52],

$$\mathcal{L} = \log p(\boldsymbol{x}) - C_{\text{IMA}}, \qquad C_{\text{IMA}} = \sum_{i=1}^{d} \log \|[J_{g^{-1}}]_i\| - \log|\det J_{g^{-1}}|. \tag{B.82}$$

This penalty is non-negative and vanishes if and only if the Jacobian is of the form QD, i.e., its columns are mutually orthogonal up to scaling. Geometrically, the columns of the Jacobian J_g correspond to the partial derivatives $\partial g/\partial s_i$. The determinant $|J_g|$ measures the volume of the d-

dimensional parallelepiped spanned by these vectors, while the product of their norms corresponds to the volume of an axis-aligned box with side lengths $\|\partial g/\partial s_i\|$. These two volumes coincide if and only if the columns are orthogonal [52]. Minimizing C_{IMA} therefore explicitly pushes the Jacobian toward an orthogonal-diagonal structure.

In our experiments, orthogonality is enforced by recovering the full Jacobian of the learned unmixing g via automatic differentiation by feeding canonical basis vectors through vector-Jacobian products using PyTorch `autograd`, and then applying the CIMA regularization term to the resulting columns. This procedure guarantees that the orthogonality condition is effectively enforced during training and that no confounding factors are introduced. While this allows us to precisely control the conditions and observe the effect of orthogonality, it requires computing the full $d \times d$ Jacobian at each training step, which scales poorly with the dimensionality of the latent space and renders direct application to high-dimensional observations such as images computationally prohibitive. In practice, efficiently enforcing orthogonal Jacobians in a scalable way is still an open research question, one that goes beyond the scope of this thesis.

Overall, we train expressive residual flows that accurately model the data distribution. In the orthogonal setting, the additional regularization enforces a QD Jacobian structure, aligning the learned transformation with the assumptions of our theoretical framework. Despite the complexity of the data-generating process, this approach yields disentangled latent representations up to coordinate-wise reparameterization.

B.3.4 Boundary Behavior and Optimization Issues

In theory, identifiability requires the learned latent distribution to exactly match the assumed prior and, in particular, to satisfy Asm.4.2 on the support of the latent variables. In practice, however, optimization and finite model capacity can prevent this condition from being met. Even when using a prior with full support, training may converge to suboptimal solutions in which the learned latent domain folds onto itself. Such configurations can achieve relatively high likelihood, remain close to the target prior, and yet violate the assumptions required by the theory. Once trapped in these configurations, continued training typically results in further local deformations of the latent space rather than convergence to the true prior, explaining most of the observed failure cases.

This also explains the observed degradation in MCC for higher latent dimensions ($d = 9$): as dimensionality increases, the C_{IMA} loss becomes harder to satisfy globally, and the optimization landscape grows more complex, making it more likely to converge to solutions where the latent domain is locally consistent with orthogonality but globally distorted. This is a limitation of the practical implementation rather than of the theory itself.

This gap highlights a fundamental difference between the theoretical setting, which assumes exact matching of the prior, and practical optimization, where approximate matching may be insufficient. To better align practice with theory, we introduce an additional mechanism to explicitly enforce correct domain geometry. The key observation is that for smooth invertible mappings, the boundary of a domain must be mapped to the boundary, and interior points must be mapped to interior points. Consequently, by enforcing correct behavior on the boundary, one indirectly constrains the entire interior of the domain.

Operationally, we identify samples lying close to the empirical boundary of the latent domain and introduce an auxiliary loss that penalizes mappings sending these points outside the expected support. This boundary aware regularization encourages the learned transformation to respect the global geometry of the domain and prevents self-intersections and holes. While this procedure requires mild prior knowledge about the data domain, it is only used to ensure that the assumptions of the identifiability theorem are met. Importantly, the method can still succeed without this regularization, but in such cases training is more prone to becoming trapped in suboptimal solutions that fall outside the theoretical regime we aim to validate empirically.

B.3.5 Experiments with Varying Latent Domains

As discussed in Sec.4.4 and further detailed in App.B.2, when the latent support is known, the generative process becomes identifiable, up to the natural rigid symmetries of that support. To support that analysis, we perform additional experiments where the latent support is known but non-cubical. Latent variables are sampled uniformly from various bounded domains and mixed using conformal Möbius transformations. For each configuration, 10 000 samples are generated. The task of the model is to recover the original latent structure solely from the observations.

For these experiments, we use Neural Spline Flows [39], which parameterize monotonic rational-quadratic spline transformations and provide high expressivity with stable training. The model consists of 12 flow layers, each implemented by a 3-layer MLP with hidden size 256, using 27 spline bins and a tail bound of 12.

To avoid optimization issues associated with sharp, bounded supports, we design a shape-aware base distribution that matches the geometry of the true latent domain. Specifically, the base density is defined as a soft-edged distribution whose density is proportional to

$$\exp\left(-\frac{\operatorname{dist}(z,\Omega)^2}{2\varepsilon^2}\right), \tag{B.83}$$

where Ω denotes the target latent domain and ε controls boundary sharpness. For large values of ε, the distribution has very smooth edges and provides gradients far outside the domain, while for small ε it approaches a uniform distribution over Ω. Points lying outside the domain experience strong gradients pulling them inward. During training, we progressively decrease ε, making the boundary increasingly sharp and, at the end of training, enforcing the target domain almost exactly. Orthogonality regularization is applied as in previous experiments.

Overall, this setup corresponds to a standard Neural Spline Flow trained with a specially designed base distribution encoding the known latent support, combined with the same orthogonality regularization used in previous experiments. This controlled setting allows us to systematically investigate the role of domain knowledge in identifiability and to empirically validate the theoretical predictions under varying and nontrivial latent geometries.

B.4 Additional Results and Experiments

In this appendix, we present additional quantitative and qualitative results complementing the experiments reported in the main chapter.

B.4.1 Quantitative Comparison

The primary goal of our experiments was to validate our theoretical results, not to establish normalizing flows as state-of-the-art disentangle-

Table B.1 Averaged disentanglement metrics across experiments.

Metric d=6, ind.	NF Unconstr.	NF Constr. (Ortho.)	VAE	β-TCVAE
MCC_S	0.7091 ± 0.0492	0.9575 ± 0.0544	0.8213 ± 0.0868	0.8880 ± 0.0759
MCC_P	0.7086 ± 0.0496	0.9510 ± 0.0545	0.8138 ± 0.0846	0.8799 ± 0.0785
MIG	0.0760 ± 0.0389	0.6138 ± 0.1429	0.2011 ± 0.0907	0.3279 ± 0.1201
SAP	0.0266 ± 0.0133	0.1356 ± 0.0236	0.0609 ± 0.0245	0.1078 ± 0.0348
DCI_D	0.1474 ± 0.0596	0.8887 ± 0.1346	0.4819 ± 0.1536	0.6766 ± 0.1665
DCI_C	0.1545 ± 0.0623	0.8898 ± 0.1326	0.5106 ± 0.1299	0.6824 ± 0.1469
DCI_I	0.9753 ± 0.0049	0.9978 ± 0.0019	0.9589 ± 0.0456	0.9734 ± 0.0335
FactorVAE	0.7195 ± 0.1279	0.9770 ± 0.0629	0.8463 ± 0.1182	0.9040 ± 0.0893

ment models. Flows were used because they satisfy the conditions of the propositions. Since VAEs already incorporate an implicit orthogonal bias (Sec.4.6), including them as a baseline was less central to our theoretical validation. Nonetheless, we provide here a comparison for $d = 6$ with independent latent variables, since the VAE prior imposes independent terms, making this the most natural setting for such a comparison.

Our initial choice of Amari distance and MCC was deliberate and aligned with prior work [52], where these metrics are commonly used and sufficient to assess whether latent factors are successfully recovered. To further strengthen our evaluation, we additionally report standard disentanglement metrics: FactorVAE [84], MIG [27], SAP [94], and DCI [41].

Results are reported in Tab.B.1. As expected from the analysis in Sec.4.6, VAEs achieve stronger disentanglement than unconstrained flows, likely due to their inherent orthogonal bias, but perform worse than constrained normalizing flows, likely due to the information preference property and the detrimental MI term. Across all additional metrics, the conclusions are consistent with those drawn from Amari distance and MCC, confirming that the learned representations are well disentangled and supporting the central claims of this work.

β-TCVAE obtains results closer to constrained NFs than standard VAEs, which is consistent with our explanation in Sec.4.6. Indeed, β-TCVAE reduces or removes the MI term while retaining (i) the total correlation penalty and (ii) a factorized posterior, thereby alleviating posterior collapse without affecting the two mechanisms responsible for disentanglement: global independence and local orthogonality. This further supports our claim that disentanglement does not arise from MI minimization, but

from the joint imposition of these two inductive biases, and that they can be made explicit and transferred beyond VAEs through orthogonality regularization. This also clarifies the comparison with unconstrained NFs: despite sharing the same total correlation pressure as VAEs, they lack the implicit orthogonality constraint induced by the factorized posterior, explaining why adding an explicit orthogonality regularization is sufficient to recover strong disentanglement.

B.4.2 Qualitative Results

Figs.B.7 and B.8 show representative reconstructions of latent variables in the case of *independent* sources, comparing unconstrained normalizing flows with their orthogonally constrained counterparts. Similarly, Figs.B.9 and B.10 report qualitative results for the *dependent* sources setting.

Across all cases, the results are consistent with the quantitative boxplot comparisons presented in Fig.4.3. Models trained with orthogonality constraints on the Jacobian are generally able to recover the underlying sources up to the expected ambiguities, whereas unconstrained models, while successfully learning the data distribution, typically fail to disentangle the true generative factors. Beyond confirming the behavior observed in the main experiments, these results demonstrate that our approach is not limited to conformal transformations or to independent source models, as considered in much of the previous literature. Instead, it successfully disentangles substantially more complex generative processes, including dependent factors and highly nonlinear, non-conformal mixings. This highlights that enforcing an orthogonal Jacobian is not merely an auxiliary regularization or architectural constraint, but rather captures a fundamental structural property of latent factors underlying identifiability in these settings.

In addition, we report results for latent domains with shapes other than the hypercube, demonstrating that our approach extends beyond standard axis-aligned supports. The experimental setup for these experiments is described in App.B.3.5, and the corresponding qualitative results are shown in FigsB.11 and B.12. As predicted by the theory (Sec.4.4 and App.B.2), the learned representations recover the sources up to the rigid symmetries of the domain, namely global rotations and reflections. This confirms that identifiability holds whenever the latent support is known, even for non-trivial geometries.

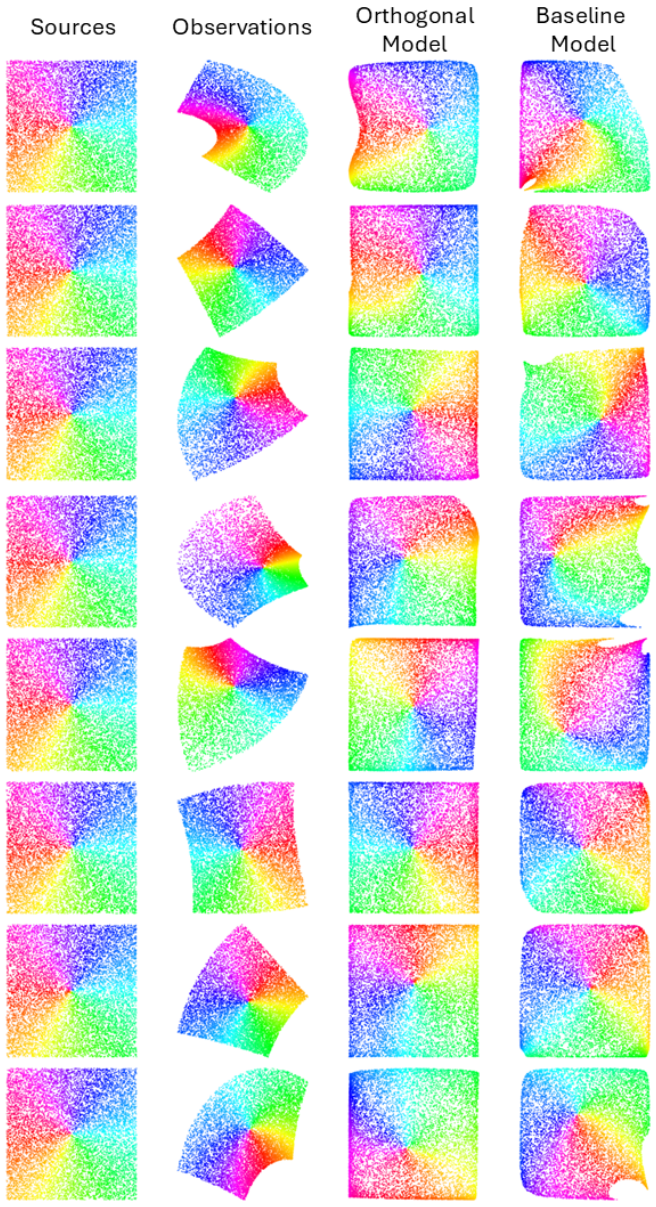


Fig. B.7 Qualitative disentanglement results for independent sources. From left to right: ground-truth sources, observed variables, latent reconstructions obtained with the orthogonally constrained model, and latent reconstructions obtained with the unconstrained model. Reconstructed latents are visualized after a sigmoid transformation for clarity.

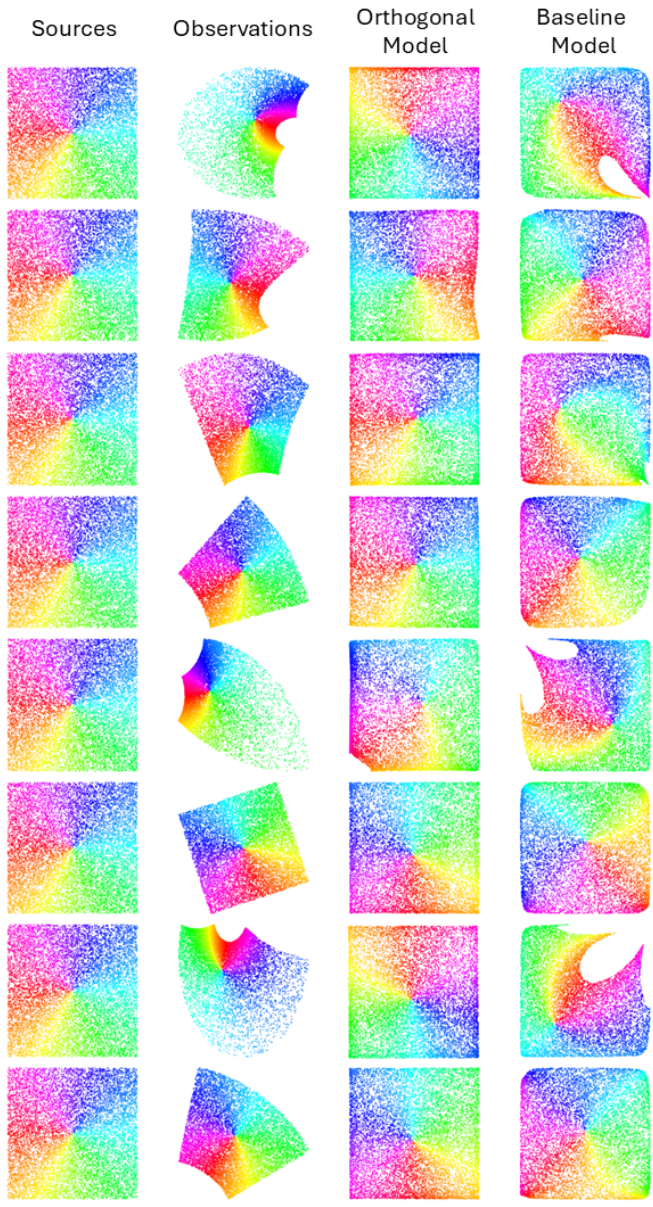


Fig. B.8 Additional qualitative results for independent sources. Columns correspond to ground-truth sources, observations, reconstructions from the constrained model, and reconstructions from the unconstrained model, respectively. Reconstructions are shown after sigmoid reparameterization.

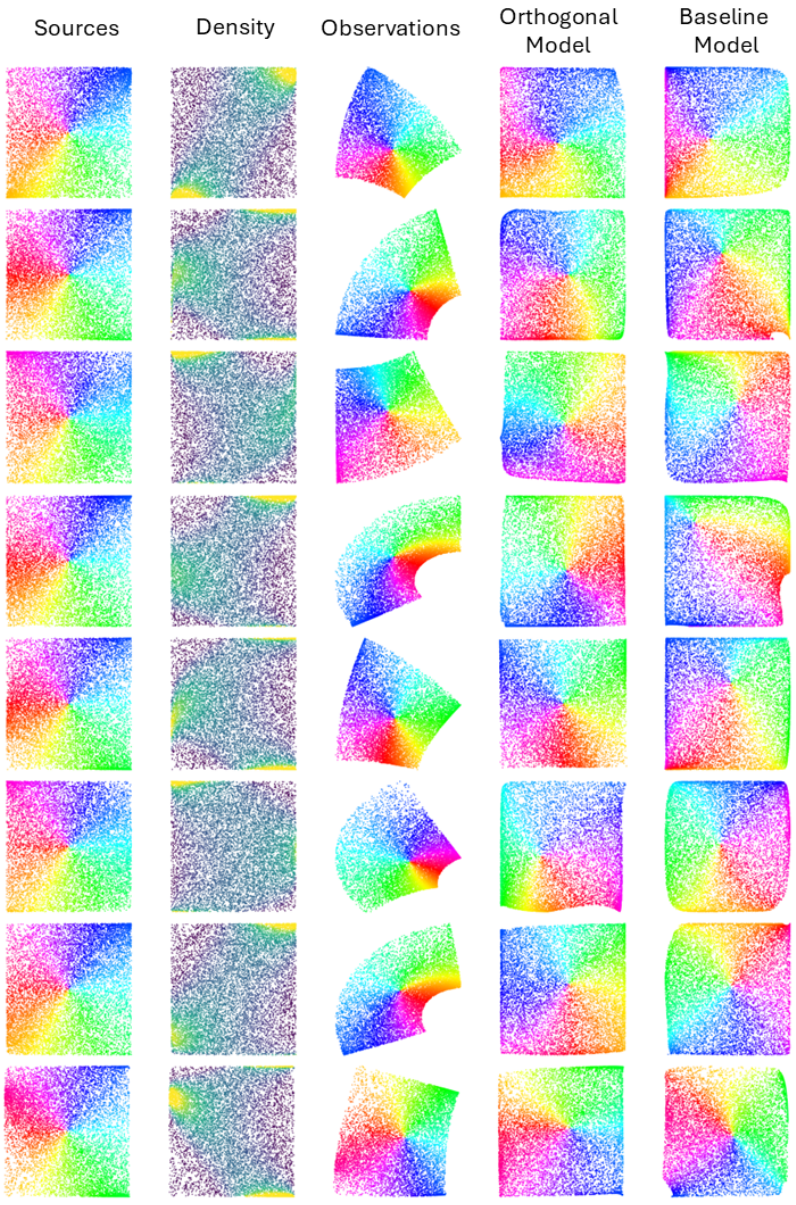


Fig. B.9 Qualitative disentanglement results for dependent sources. From left to right: ground-truth sources, source density, observed variables, latent reconstructions obtained with the orthogonally constrained model, and latent reconstructions obtained with the unconstrained model. Reconstructions are visualized after sigmoid transformation.

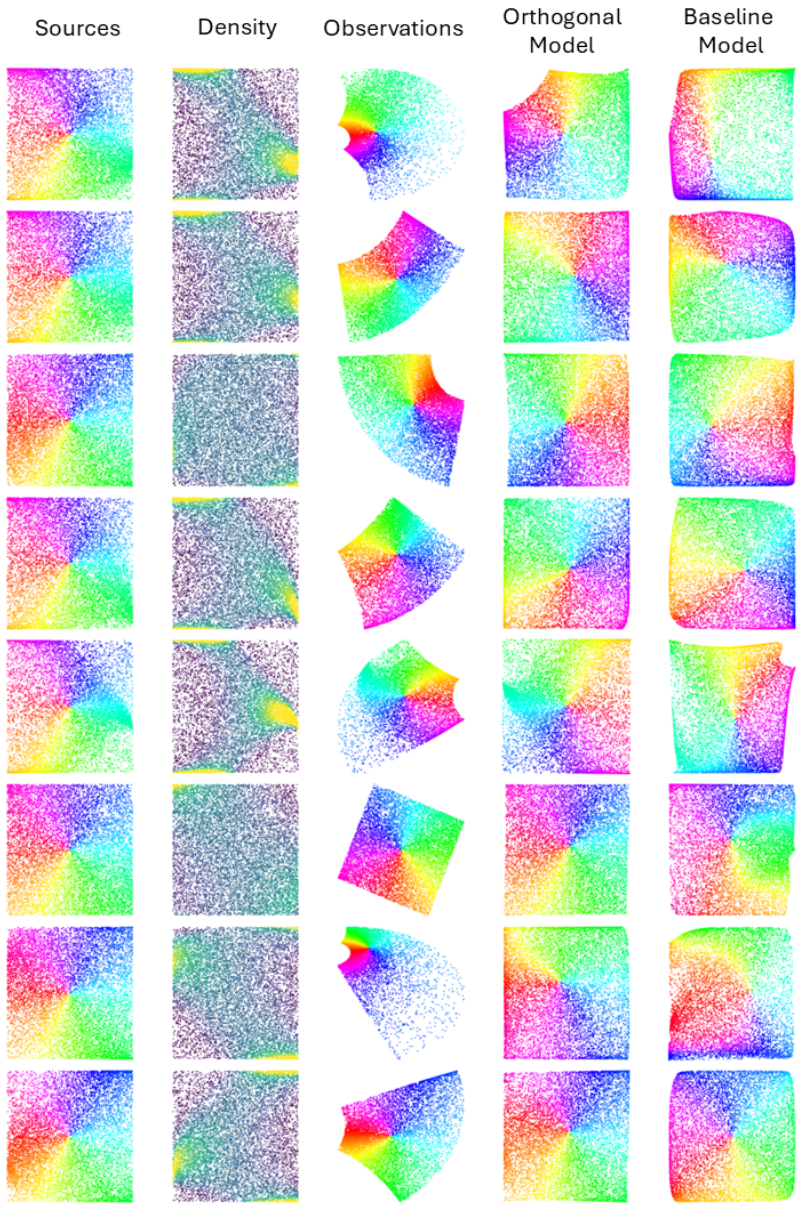


Fig. B.10 Additional qualitative results for dependent sources. Columns show ground-truth sources, source density, observations, and latent reconstructions from the constrained and unconstrained models, respectively.

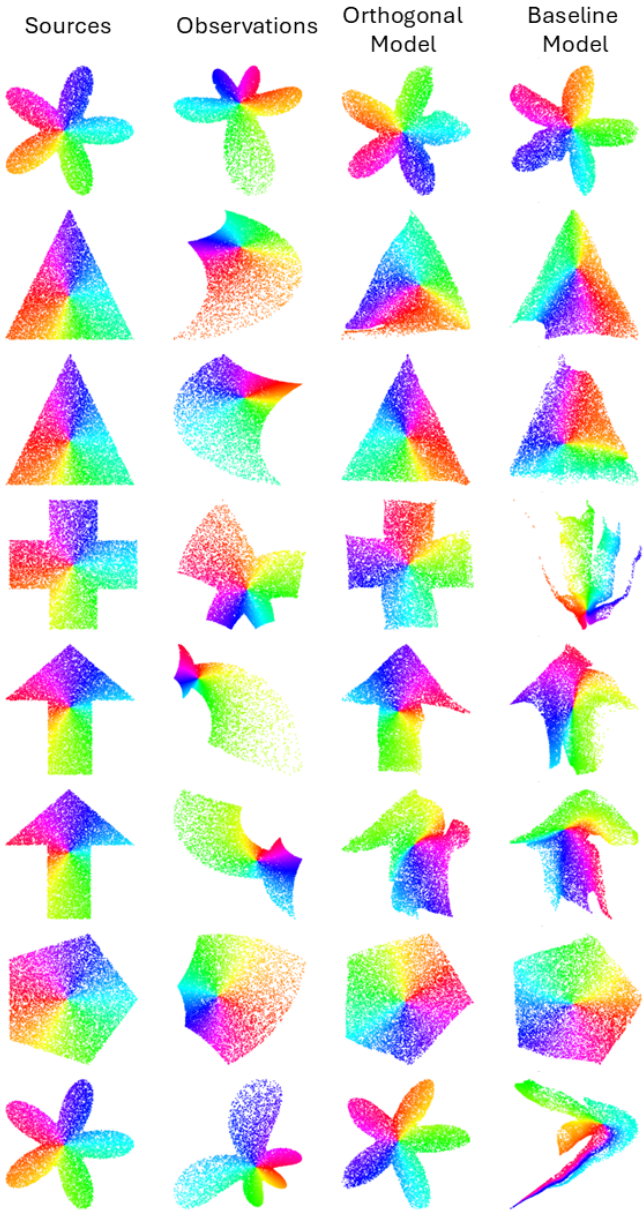


Fig. B.11 Qualitative disentanglement results for non-cubical latent domains. From left to right: ground-truth sources sampled from a structured domain, observations, reconstructions obtained with the orthogonally constrained model, and reconstructions obtained with the unconstrained model.

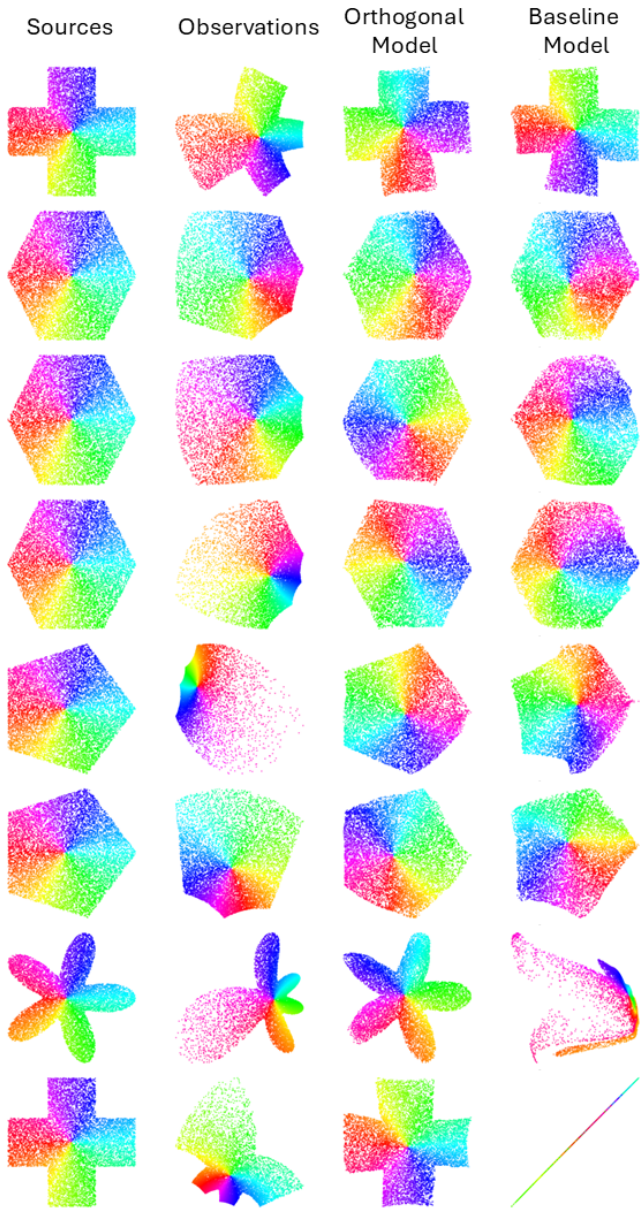


Fig. B.12 Additional results for non-cubical latent domains. The constrained model successfully recovers the sources up to rigid transformations of the domain, while the unconstrained model fails to preserve the latent geometry.

C

Supplementary Material for Chapter 5

C.1 Dataset Descriptions

We evaluate our model on six video datasets spanning a diverse range of visual content, motion complexity, and label structure. All datasets are pre-processed to a common format: pixel values are normalized to $[-1, 1]$ on a per-sequence basis, and sequences are stored as float tensors in $\mathbb{R}^{T \times C \times H \times W}$.

MUG [3] (*MUG Facial Expression Database*) is a facial expression dataset comprising video sequences of 52 subjects performing six different expressions: anger, disgust, fear, happiness, sadness, and surprise. Following [141], we apply Haar Cascade face detection to crop the face region and resize each frame to 64×64. Each take is converted into four temporally-shifted length-15 clips by equally spacing 15 frames from shifted starting offsets that together span the full source video, yielding 3636 sequences in total. The dynamic label is the emotion class and the static label is the subject identity.

MHAD [25] (*UTD Multimodal Human Action Dataset*) is a real-world action recognition dataset comprising video sequences of 8 subjects (4 female, 4 male) performing 27 different actions such as waving, sitting, throwing, and walking, captured with a Microsoft Kinect camera at 640×480

resolution. Each action is repeated four times per subject. We apply a center crop of 240×240 and resize to 64×64. As for MUG, each source video yields four temporally-shifted length-15 clips, producing 3444 sequences in total. The dynamic label is the action class and the static label is the subject identity.

KTH [135] (*KTH Action Recognition dataset*) is a grayscale human action recognition dataset containing videos of 25 subjects performing 6 actions (boxing, handclapping, handwaving, jogging, running, and walking) recorded at 160×120 resolution and 25 fps under four different recording conditions. Videos are center-cropped to 120×120 and resized to 64×64. Each video is segmented into non-overlapping 1.5-second blocks of 37 frames, from which 15 frames are uniformly subsampled (approximately 10 fps), yielding 7534 sequences in total. Labels include the action class, subject identity, and recording condition.

Moving MNIST [142] is a synthetic grayscale benchmark in which two MNIST digits move independently with constant velocity and bounce off image boundaries. We use the standard test set of 10 000 sequences, each consisting of 20 frames at 64×64 resolution. Each sequence carries a static label corresponding to the digit identity. As no dynamic label is available, Moving MNIST is used to evaluate static representation quality and density estimation. The single-channel sequences are rescaled to $[-1, 1]$.

LPCSprites [128] (*Universal LPC Character dataset*) is a synthetic dataset of animated sprite characters, where each sequence depicts a character performing one of 9 actions across 8 frames at 64×64 resolution in color. Characters are procedurally generated with independently varying attributes including body color, shirt type and color, pant type and color, hair style and color, and viewing orientation, yielding a rich multi-label structure. The dataset is provided with a fixed train/test split and the sprite tensors are rescaled from $[0, 1]$ to $[-1, 1]$. The dynamic label is the action class and the static label is the subject identity (defined by the combination of visual attributes).

BAIR [42] (*Berkeley AI Research Robot Pushing Dataset*) is a real-world dataset of a robotic arm pushing objects on a table, captured at 64×64 resolution in color. The dataset contains a large variety of object configurations and motion patterns, with no semantic labels. Each source sequence consists of 30 frames; we split each into two non-overlapping 15-frame clips, yielding 87 040 sequences in total. As no class annotations are available, BAIR is evaluated through density estimation and label-free metrics only.

C.2 Evaluation Metrics and Additional Discussions

We evaluate the proposed model along two complementary axes: the informational content of the global latent variable $\boldsymbol{s}$, and its effect on density estimation. The former is assessed through a combination of label-based and label-free metrics; the latter through bits-per-dimension (BPD) measurements.

C.2.1 Label-Based Metrics

For datasets equipped with sequence-level annotations, we exploit the available labels to assess the discriminative structure of the learned representations. We evaluate two representations: the global latent $\boldsymbol{s} \in \mathcal{S}$, obtained as the output of the sequence encoder; and the collection of frame-wise residuals $\boldsymbol{\tau}_{1:T} = (\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_T)$, obtained as the outputs of the conditional normalizing flow. We report metrics for both dynamic labels (e.g., action class) and static labels (e.g., subject identity) where available.

Linear and k-NN accuracy. We train a linear classifier [121] and a k-nearest-neighbor (k-NN) classifier [31] on top of each representation and evaluate accuracy via 5-fold stratified cross-validation. To prevent memorization and ensure that classifiers rely on representation geometry rather than dataset size, we use a stratified subsample of at most 250 training examples. The linear classifier consists of standardization, PCA reduction to 32 components, and an ℓ_2-regularized logistic regression. We report k-NN accuracy for $k \in \{1, 3, 5\}$ and report the best result across all values of k.

Fisher discriminant ratio. For each label key k, we measure the separability of the representation with respect to label k while controlling for all other label dimensions, so as to isolate the contribution of each factor independently. For a query sample i carrying label value v_k for key k and a fixed combination of values v_{-k} on all other keys, we define the set of intra-class partners $\mathcal{I}_i = \{j \neq i : \ell_k^j = v_k,\ \ell_{-k}^j = v_{-k}\}$ and inter-class partners $\mathcal{O}_i = \{j : \ell_k^j \neq v_k,\ \ell_{-k}^j = v_{-k}\}$. Queries for which either set is empty are discarded. The Fisher discriminant ratio [43] for label key k is then defined as

$$R^{(k)}_{\text{Fisher}} = \frac{\frac{1}{|\mathcal{Q}|}\sum_{i\in\mathcal{Q}} \mathbb{E}_{j\sim\mathcal{O}_i}\left[\|f_i - f_j\|_2^2\right]}{\frac{1}{|\mathcal{Q}|}\sum_{i\in\mathcal{Q}} \mathbb{E}_{j\sim\mathcal{I}_i}\left[\|f_i - f_j\|_2^2\right]}, \tag{C.1}$$

where f_i denotes the representation of sample i and $\mathcal{Q}$ is the set of valid query samples. A ratio of $R^{(k)}_{\text{Fisher}} = 1$ indicates that inter-class and intra-class distances are equal, i.e., the representation carries no discriminative information about label k. Values above 1 indicate increasing separability. We report the normalized ratio

$$\tilde{R}^{(k)}_{\text{Fisher}} = \frac{R^{(k)}_{\text{Fisher}}}{(R^{(k)}_{\text{Fisher}} + 1)} \in [0, 1], \tag{C.2}$$

where the normalization maps $R^{(k)}_{\text{Fisher}} = 1$ to $\tilde{R} = 0.5$ (no separability), values approaching 0 indicate poor structure (intra-class distances dominate), and values approaching 1 correspond to perfect separation ($R^{(k)}_{\text{Fisher}} \to \infty$). Due to the nonlinearity of this mapping, even modest raw ratios correspond to noticeably high normalized values: for instance, $R^{(k)}_{\text{Fisher}} = 3$ yields $\tilde{R} = 0.75$, and $R^{(k)}_{\text{Fisher}} = 9$ yields $\tilde{R} = 0.9$.

Silhouette score. We additionally assess the cluster structure of each representation via the silhouette score [132]. For each label key k, we fix all other label dimensions and compute, within each group of samples sharing the same context v_{-k}, the silhouette coefficient of each sample i:

$$\text{sil}(i) = \frac{b(i) - a(i)}{\max(a(i),\, b(i))}, \tag{C.3}$$

where $a(i) = \frac{1}{|\mathcal{I}_i|}\sum_{j\in\mathcal{I}_i} d(f_i, f_j)$ is the mean distance to all other samples in the same class, and $b(i) = \min_{c\neq v_k} \frac{1}{|\mathcal{O}_i^c|}\sum_{j\in\mathcal{O}_i^c} d(f_i, f_j)$ is the mean distance to the samples of the nearest other class c, i.e., the class whose mean distance to i is smallest. The overall silhouette score is the mean of $\text{sil}(i)$ over all valid samples. The silhouette score lies in $[-1, 1]$: a value near $+1$ indicates that samples are tightly clustered around their own class and well separated from the nearest other class; a value near 0 indicates that samples lie close to the boundary between their class and the nearest neighbor class; and negative values indicate that samples are on average closer to the nearest other class than to their own, suggesting a poorly structured representation with respect to label k.

Label transfer. To assess whether the global latent $\boldsymbol{s}$ encodes semantically stable content, we evaluate label consistency under sampling from the conditional decoder. Given a sequence with label y, we fix its global latent $\boldsymbol{s}$ and generate $K = 15$ pseudo-frames by sampling from $p_\theta(\boldsymbol{x} \mid \boldsymbol{s})$. We re-encode these samples through the sequence encoder to obtain a new latent $\boldsymbol{s}'$, which is then classified using the linear classifier trained on $\boldsymbol{s}$. We report the fraction of generated samples whose predicted label matches the original label predicted on $\boldsymbol{s}$, thereby measuring whether the semantic content encoded in $\boldsymbol{s}$ is faithfully preserved through the generation and re-encoding pipeline. A high score indicates that sampling $\boldsymbol{\tau}_{1:T}$ from the conditional decoder given a fixed $\boldsymbol{s}$ does not alter the predicted label of the sequence, confirming that the residual latents carry only limited information beyond what is already encoded in $\boldsymbol{s}$.

Remark : Since $\boldsymbol{s}$ is a single fixed-size vector, classifiers are trained directly on $\boldsymbol{s}$ and pairwise distances are computed as standard Euclidean distances $\|\boldsymbol{s} - \boldsymbol{s}'\|_2$. For the residual latents, however, each sequence is associated with a collection $\boldsymbol{\tau}_{1:T}$. To enable classification, we consider two strategies: training the classifier on the temporal average $\bar{\boldsymbol{\tau}} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\tau}_t$, and training on individual $\boldsymbol{\tau}_t$ values with prediction obtained by majority vote across frames. We report the best result across both strategies. For pairwise distances between two sequences $\boldsymbol{\tau}_{1:T}$ and $\boldsymbol{\tau}'_{1:T}$, we compute the minimum-cost assignment distance, defined as $d(\boldsymbol{\tau}_{1:T}, \boldsymbol{\tau}'_{1:T}) = \min_\sigma \sum_{t=1}^{T} \|\boldsymbol{\tau}_t - \boldsymbol{\tau}'_{\sigma(t)}\|_2$, where σ ranges over all permutations of $\{1, \ldots, T\}$. This assignment-based distance is used in all metric computations involving $\boldsymbol{\tau}_{1:T}$.

C.2.2 Label-Free Metrics

For all datasets, including those without semantic annotations, we assess the informative content of $\boldsymbol{s}$ through metrics that do not rely on external supervision.

Discriminative score. Given a sequence $\boldsymbol{x}_{1:T}$ and its associated global latent $\boldsymbol{s}$, we define the discriminative score as

$$d_{\text{score}} = \mathbb{E}_{p(\boldsymbol{x}_{1:T})} \frac{p_\theta(\boldsymbol{x}_{1:T} \mid \boldsymbol{s})}{p_\theta(\boldsymbol{x}_{1:T} \mid \boldsymbol{s}) + \sum_{i=1}^{K} p_\theta(\boldsymbol{x}_{1:T} \mid \boldsymbol{s}'_i)}, \tag{C.4}$$

where $\{s'_i\}_{i=1}^K$ are global latents inferred from K other sequences drawn from the dataset. This score takes values in $[0,1]$ and can be interpreted as the probability that the model assigns higher likelihood to the true conditioning latent s than to each of K randomly drawn alternatives. A score of $\frac{1}{K+1}$ corresponds to chance level, while a score close to 1 indicates that the model strongly and consistently prefers the correct latent. This metric is directly inspired by regularization scores used to train models in [7, 69, 68].

This score is closely related to the InfoNCE objective [114]. Recall that InfoNCE with K negatives and $f(x_{1:T}, s) = p_\theta(x_{1:T} \mid s)$ is defined as

$$\mathcal{L}_{\text{NCE}} = \mathbb{E}\left[\log \frac{p_\theta(x_{1:T} \mid s)}{p_\theta(x_{1:T} \mid s) + \sum_{i=1}^K p_\theta(x_{1:T} \mid s'_i)}\right], \tag{C.5}$$

which is precisely $\mathbb{E}[\log d_{\text{score}}^{(i)}]$, the expectation of the log of the per-sample ratio. Our d_{score} moves the expectation inside (before the log), yielding a quantity in $[0,1]$ that is easier to interpret directly as a probability. By Jensen's inequality [80], $\log d_{\text{score}} \geq \mathcal{L}_{\text{NCE}}$, and $\log d_{\text{score}}$ itself provides a lower bound on $I(x_{1:T}; s)$, as established in the proposition below.

Proposition C.1. $\log(K+1) + \log d_{\text{score}} \leq I(x_{1:T}; s)$.

Proof. Let $s'_1, \ldots, s'_K$ be latent variables inferred from K sequences drawn independently from the data distribution $p(x_{1:T})$. We have

$$\begin{aligned} d_{\text{score}} &= \mathbb{E}_{p(x_{1:T})} \frac{p_\theta(x_{1:T} \mid s)}{p_\theta(x_{1:T} \mid s) + \sum_{i=1}^K p_\theta(x_{1:T} \mid s'_i)} \\ &\approx \mathbb{E}_{p(x_{1:T})} \frac{1}{1 + \frac{p(x_{1:T})}{p_\theta(x_{1:T}|s)} \cdot K \cdot \mathbb{E}_{s'}\left[\frac{p_\theta(x_{1:T}|s')}{p(x_{1:T})}\right]} \\ &= \mathbb{E}_{p(x_{1:T})} \frac{1}{1 + K \cdot \frac{p(x_{1:T})}{p_\theta(x_{1:T}|s)}}, \end{aligned} \tag{C.6}$$

where the approximation replaces the finite sum over K negatives by its expectation under the marginal $p(s)$, and the last equality uses the marginalization identity $\mathbb{E}_{s' \sim p(s)}[p_\theta(x_{1:T} \mid s')] = p(x_{1:T})$. Taking the logarithm and applying Jensen's inequality,

$$\log d_{\text{score}} \lesssim \mathbb{E}_{p(x_{1:T})} \log \frac{1}{1 + K \cdot \frac{p(x_{1:T})}{p_\theta(x_{1:T}|s)}}$$

$$
\begin{aligned}
&= -\mathbb{E}_{p(\boldsymbol{x}_{1:T})} \log \left(1 + K \cdot \frac{p(\boldsymbol{x}_{1:T})}{p_\theta(\boldsymbol{x}_{1:T} \mid \boldsymbol{s})}\right) \\
&= -\mathbb{E}_{p(\boldsymbol{x}_{1:T})} \log \left(\frac{1}{K+1} + \frac{K}{K+1} \cdot \frac{p(\boldsymbol{x}_{1:T})}{p_\theta(\boldsymbol{x}_{1:T} \mid \boldsymbol{s})}\right) - \log(K+1) \\
&\leq -\frac{1}{K+1} \log 1 - \frac{K}{K+1} \mathbb{E}_{p(\boldsymbol{x}_{1:T})} \log \frac{p(\boldsymbol{x}_{1:T})}{p_\theta(\boldsymbol{x}_{1:T} \mid \boldsymbol{s})} - \log(K+1) \\
&= \frac{K}{K+1} \mathbb{E}_{p(\boldsymbol{x}_{1:T})} \log \frac{p_\theta(\boldsymbol{x}_{1:T} \mid \boldsymbol{s})}{p(\boldsymbol{x}_{1:T})} - \log(K+1) \\
&\approx I(\boldsymbol{x}_{1:T}; \boldsymbol{s}) - \log(K+1), \qquad \text{(C.7)}
\end{aligned}
$$

where the inequality applies Jensen's inequality to the concave log of the convex combination $\frac{1}{K+1} \cdot 1 + \frac{K}{K+1} \cdot \frac{p(\boldsymbol{x}_{1:T})}{p_\theta(\boldsymbol{x}_{1:T} \mid \boldsymbol{s})}$, and the last approximation uses $\frac{K}{K+1} \approx 1$ for large K. Rearranging gives $I(\boldsymbol{x}_{1:T}; \boldsymbol{s}) \geq \log(K+1) + \log d_{\text{score}}$. □

In our setting, the exact log-likelihood $\log p_\theta(\boldsymbol{x}_{1:T} \mid \boldsymbol{s})$ is directly available from the normalizing flow, and d_{score} is computed using this exact likelihood. To further validate the discriminative content of $\boldsymbol{s}$ through an independent measure that does not rely on likelihood values, we also compute a distance-based variant. Given a sequence, we generate a pool of $M = 200$ samples from the conditional decoder under the correct $\boldsymbol{s}$ and under a randomly drawn $\boldsymbol{s}'$, and compute the mean nearest-neighbor distance from the true frames to each pool. The distance-based discriminative score is then defined analogously to d_{score}, replacing the likelihood ratio by the inverse of these nearest-neighbor distances. Both variants are reported in Tab.5.2 as d_{score} and dist-d_{score} respectively.

Jacobian sensitivity ratio. To quantify the relative contribution of the global latent $\boldsymbol{s}$ and the residual latent $\boldsymbol{\tau}_t$ to the variability of the decoder output, we estimate the Frobenius norms of the Jacobians of the decoder h_θ with respect to each input. Denoting $J_\tau = \partial h / \partial \boldsymbol{\tau}$ and $J_{\boldsymbol{s}} = \partial h / \partial \boldsymbol{s}$, we define the ratio

$$
R_\tau = \frac{\mathbb{E}[\|J_\tau\|_F^2]}{\mathbb{E}[\|J_\tau\|_F^2] + \mathbb{E}[\|J_{\boldsymbol{s}}\|_F^2]}, \qquad \text{(C.8)}
$$

which measures the fraction of output variability attributable to the residual latent. A low value of R_τ indicates that most variability in the decoder output is driven by $\boldsymbol{s}$. Both norms are estimated via a randomized finite-difference Hutchinson estimator [72] using $K = 20$ random directions per sample, averaged over 64 sequences.

C.2.3 Density Estimation

Bits-per-dimension. We evaluate the quality of the density model in terms of bits-per-dimension (BPD), a standard metric for likelihood-based generative models [36, 115]. Since the model operates on discrete quantized latent representations, uniform dequantization noise is added during training following standard practice [36, 64], transforming the discrete data distribution into a continuous one. This allows the model to learn a valid density by modeling the noise-augmented distribution. BPD is then reported with respect to the original discrete inputs $\tilde{\boldsymbol{x}}$ via

$$\tilde{p}(\tilde{\boldsymbol{x}}) = \int_{\boldsymbol{\epsilon} \in [0,\delta]^D} p_\theta(\tilde{\boldsymbol{x}} + \boldsymbol{\epsilon})\, \mathrm{d}\boldsymbol{\epsilon}, \tag{C.9}$$

where δ is the dequantization bin width and D is the dimension of the latent representation. In practice, following [36, 175], this integral is estimated by a single-sample Monte Carlo approximation, i.e., a single dequantization noise draw per training step, which is standard practice and provides an unbiased estimator of the log-likelihood [130, 115].

BPD is computed as the empirical average of the negative log-likelihood over the dataset, normalized by the number of dimensions:

$$\mathrm{BPD} = -\frac{1}{D \log 2} \cdot \frac{1}{N} \sum_{n=1}^{N} \log p_\theta(\boldsymbol{x}_n), \tag{C.10}$$

where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ are i.i.d. samples from the data distribution p_d. This empirical average is a standard Monte Carlo estimate of the expected negative log-likelihood $\mathbb{E}_{p_d}[-\log p_\theta(\boldsymbol{x})]$, which converges to the true expectation as $N \to \infty$. This evaluation strategy is the same principle underlying maximum likelihood training and is standard in the literature [130, 115].

We consider two complementary settings. In the *frame-independent* setting, frames are modeled without temporal dependencies, corresponding to an intra-frame encoding regime. In the *autoregressive* setting, a learned prior is introduced over the residual latents $\boldsymbol{\tau}_t$, allowing the model to exploit temporal dependencies across frames.

For the unconditional model, the density is estimated directly via the normalizing flow, yielding $\log p_\theta(\boldsymbol{x}_{1:T})$ as an exact log-likelihood, which is then reported in BPD. For the conditional model, the framework is analogous to a VAE with a conditional normalizing flow decoder as described in Sec.5.4: the marginal log-likelihood $\log p_\theta(\boldsymbol{x}_{1:T})$ is in general intractable

due to the integration over $\boldsymbol{s}$. We instead report the joint log-likelihood $\log p_\theta(\boldsymbol{x}_{1:T}, \boldsymbol{s})$, which decomposes as

$$\log p_\theta(\boldsymbol{x}_{1:T}, \boldsymbol{s}) = \log p_\theta(\boldsymbol{s}) + \log p_\theta(\boldsymbol{x}_{1:T} \mid \boldsymbol{s}), \tag{C.11}$$

where the conditional term $\log p_\theta(\boldsymbol{x}_{1:T} \mid \boldsymbol{s})$ is the exact log-likelihood from the conditional normalizing flow, and $\log p_\theta(\boldsymbol{s})$ is the log-likelihood under a learned prior on $\boldsymbol{s}$. This joint likelihood therefore accounts for two distinct coding costs: the cost of encoding $\boldsymbol{s}$ itself through $\log p_\theta(\boldsymbol{s})$, and the cost of encoding the sequence conditionally on $\boldsymbol{s}$ through $\log p_\theta(\boldsymbol{x}_{1:T} \mid \boldsymbol{s})$, providing a complete and conservative accounting of the total description length.

As described in Sec.5.4, $\boldsymbol{s}$ is constrained to lie in the same bounded range $[-1, 1]$ as the input frames, and the same fixed-variance dequantization noise is added to $\boldsymbol{s}$ to ensure consistency with the continuous modeling framework. Furthermore, for density estimation, $\boldsymbol{s}$ is treated as a quantized latent representation with 256 discrete levels and bin width $\delta = 2/256$, consistent with the autoencoder latents (App.C.3), using a straight-through estimator (STE) during training. This ensures that $\log p_\theta(\boldsymbol{s})$ is a well-defined probability mass over a discrete support, and that the joint log-likelihood $\log p_\theta(\boldsymbol{x}_{1:T}, \boldsymbol{s})$ is a proper quantity in bits, amenable to interpretation as a BPD measure.

We now show that this quantity coincides with the ELBO on the marginal $\log p_\theta(\boldsymbol{x}_{1:T})$. The ELBO is defined as:

$$\log p_\theta(\boldsymbol{x}_{1:T}) \geq \mathbb{E}_{q_\phi(\boldsymbol{s}|\boldsymbol{x}_{1:T})} \left[\log p_\theta(\boldsymbol{x}_{1:T} \mid \boldsymbol{s})\right] - \mathrm{KL}\left(q_\phi(\boldsymbol{s} \mid \boldsymbol{x}_{1:T}) \,\|\, p(\boldsymbol{s})\right). \tag{C.12}$$

In our model, $\boldsymbol{s}$ is quantized and the encoder is deterministic up to fixed additive noise: the posterior $q_\phi(\boldsymbol{s} \mid \boldsymbol{x}_{1:T})$ is therefore a point mass concentrated on a single quantized value $\boldsymbol{s}^*$, i.e. $q_\phi(\boldsymbol{s} \mid \boldsymbol{x}_{1:T}) = \delta_{\boldsymbol{s}^*}$. The KL divergence then reduces exactly to:

$$\begin{aligned} \mathrm{KL}\left(q_\phi(\boldsymbol{s} \mid \boldsymbol{x}_{1:T}) \,\|\, p(\boldsymbol{s})\right) &= \sum_{\boldsymbol{s}} q_\phi(\boldsymbol{s} \mid \boldsymbol{x}_{1:T}) \log \frac{q_\phi(\boldsymbol{s} \mid \boldsymbol{x}_{1:T})}{p(\boldsymbol{s})} \\ &= \log \frac{1}{p(\boldsymbol{s}^*)} \\ &= -\log p(\boldsymbol{s}^*), \end{aligned} \tag{C.13}$$

since the point mass assigns probability one to $\boldsymbol{s}^*$ and zero elsewhere, so

the entropy of $q_\phi(\boldsymbol{s} \mid \boldsymbol{x}_{1:T})$ vanishes. Substituting into the ELBO:

$$\text{ELBO} = \log p_\theta(\boldsymbol{x}_{1:T} \mid \boldsymbol{s}^*) + \log p(\boldsymbol{s}^*) = \log p_\theta(\boldsymbol{x}_{1:T}, \boldsymbol{s}^*), \tag{C.14}$$

which is exactly the joint log-likelihood. The reported joint likelihood therefore coincides with the ELBO, and constitutes simultaneously: (i) an *upper bound on the conditional BPD* of the true marginal $p_\theta(\boldsymbol{x}_{1:T})$, since the ELBO is a lower bound on the marginal log-likelihood; and (ii) a *complete accounting of the total coding cost*, since it explicitly includes both the cost of transmitting $\boldsymbol{s}$ under the learned prior $p_\theta(\boldsymbol{s})$ and the cost of encoding the sequence given $\boldsymbol{s}$. This means the true marginal BPD of the conditional model would be even lower than reported, so the BPD improvements shown in Tab.5.3 are conservative estimates of the true gain from conditioning on $\boldsymbol{s}$.

Information-theoretic interpretation of the BPD gap. Beyond being a measure of model fit, the gap between unconditional and conditional BPD admits a direct information-theoretic interpretation that connects density estimation to the mutual information between the sequence and its global latent representation.

Since frames are modeled independently in both the unconditional and conditional settings, the unconditional BPD estimates $\mathbb{E}_{p_d}\left[-\sum_{t=1}^{T} \log p_\theta(\boldsymbol{x}_t)\right]$, which decomposes as

$$\mathbb{E}_{p_d}\left[-\sum_{t=1}^{T} \log p_\theta(\boldsymbol{x}_t)\right] = \sum_{t=1}^{T} H(\boldsymbol{x}_t) + \sum_{t=1}^{T} \mathrm{KL}(p_d(\boldsymbol{x}_t) \| p_\theta(\boldsymbol{x}_t)), \tag{C.15}$$

where $H(\boldsymbol{x}_t)$ is the marginal entropy of frame t and the KL terms capture the per-frame modeling errors of the unconditional model. The conditional BPD estimates $\mathbb{E}_{p_d}\left[-\sum_{t=1}^{T} \log p_\theta(\boldsymbol{x}_t \mid \boldsymbol{s}) - \log p_\theta(\boldsymbol{s})\right]$, which decomposes as

$$\begin{aligned}\mathbb{E}_{p_d}\left[-\sum_{t=1}^{T} \log p_\theta(\boldsymbol{x}_t \mid \boldsymbol{s}) - \log p_\theta(\boldsymbol{s})\right] &= \sum_{t=1}^{T} H(\boldsymbol{x}_t) - \sum_{t=1}^{T} I(\boldsymbol{x}_t; \boldsymbol{s}) + H(\boldsymbol{s}) \\ &\quad + \sum_{t=1}^{T} \mathrm{KL}(p_d(\boldsymbol{x}_t \mid \boldsymbol{s}) \| p_\theta(\boldsymbol{x}_t \mid \boldsymbol{s})) \\ &\quad + \mathrm{KL}(p_d(\boldsymbol{s}) \| p_\theta(\boldsymbol{s})),\end{aligned} \tag{C.16}$$

where $H(\boldsymbol{s})$ is the entropy of the prior on $\boldsymbol{s}$, $\sum_{t=1}^{T} \mathrm{KL}(p_d(\boldsymbol{x}_t \mid \boldsymbol{s}) \| p_\theta(\boldsymbol{x}_t \mid \boldsymbol{s}))$ are the per-frame conditional modeling errors, and $\mathrm{KL}(p_d(\boldsymbol{s}) \| p_\theta(\boldsymbol{s}))$ is the error of the prior on $\boldsymbol{s}$. The $\sum_{t=1}^{T} H(\boldsymbol{x}_t)$ terms cancel in the gap, which therefore satisfies

$$\begin{aligned}\Delta\mathrm{BPD} = \sum_{t=1}^{T} I(\boldsymbol{x}_t; \boldsymbol{s}) - H(\boldsymbol{s}) - \sum_{t=1}^{T} \mathrm{KL}(p_d(\boldsymbol{x}_t \mid \boldsymbol{s}) \| p_\theta(\boldsymbol{x}_t \mid \boldsymbol{s})) \\ - \mathrm{KL}(p_d(\boldsymbol{s}) \| p_\theta(\boldsymbol{s})) + \sum_{t=1}^{T} \mathrm{KL}(p_d(\boldsymbol{x}_t) \| p_\theta(\boldsymbol{x}_t)).\end{aligned} \tag{C.17}$$

Since both the conditional and unconditional models share the same normalizing flow architecture, their respective per-frame modeling errors are expected to be of comparable magnitude and approximately cancel. Furthermore, $\mathrm{KL}(p_d(\boldsymbol{s}) \| p_\theta(\boldsymbol{s}))$ is expected to be small when the prior on $\boldsymbol{s}$ is well fitted. Under these assumptions, the gap reduces to

$$\Delta\mathrm{BPD} \approx \sum_{t=1}^{T} I(\boldsymbol{x}_t; \boldsymbol{s}) - H(\boldsymbol{s}), \tag{C.18}$$

so that the observed BPD improvement provides a direct proxy for the total mutual information accumulated across all frames with respect to the global latent $\boldsymbol{s}$, up to the entropy cost of encoding $\boldsymbol{s}$ itself.

This quantity can be large when $\boldsymbol{s}$ simultaneously captures information shared across many frames: unlike a single-frame model where the gain would be bounded by the entropy of a single frame, a global latent inferred from the full sequence can accumulate mutual information across all T frames, yielding gains that grow with sequence length. In this regime, the unconditional model must independently model the full marginal uncertainty of each frame, while the conditional model encodes the shared structure (both static content and global dynamic information) once in $\boldsymbol{s}$, reducing the residual uncertainty at each frame to its minimum. This provides an additional validation of the theoretical motivation developed in Sec.5.3: the gain from conditioning on $\boldsymbol{s}$ is large precisely when $\boldsymbol{s}$ concentrates maximal information about the sequence.

Implications for compression. Video data represents a dominant fraction of global internet traffic [8]. Consequently, even modest improvements in coding efficiency can have substantial environmental and economic impact. Under Shannon's source coding theorem, the expected code length of

a source is bounded below by its entropy, so any reduction in the modeled entropy of the data distribution translates directly into improved coding efficiency. From Eq.(C.18), the BPD gap approximates $\sum_t I(\boldsymbol{x}_t; \boldsymbol{s}) - H(\boldsymbol{s})$, which is positive whenever the mutual information accumulated across frames exceeds the cost of encoding $\boldsymbol{s}$ itself. In our framework, $\boldsymbol{s}$ acts as a sequence-level side-information that captures structure shared across all frames, including scene content, object identities, and global motion patterns. Encoding $\boldsymbol{s}$ once at the sequence level and conditioning every frame upon it therefore reduces temporal redundancy in the factorization of the data distribution, in a manner conceptually analogous to intra-frame reference coding in classical video codecs, but learned end-to-end and without explicit motion estimation. In the autoregressive setting, the global latent $\boldsymbol{s}$ further provides long-range context that complements the local information captured by past frames, reducing the residual uncertainty that the temporal prior must model.

C.3 Architecture and Hyperparameters

Training high-capacity NFs directly in image space is computationally expensive. To address this, all components of the model operate in the compressed latent space of a pretrained autoencoder rather than in pixel space. We first pretrain an autoencoder to reduce the dimensionality of the frames and obtain discrete, quantized latent representations. The sequence encoder and conditional NF are then trained on these latent codes rather than raw pixels. This allows efficient training while preserving the probabilistic decomposition of sequence information into the global component $\boldsymbol{s}$ and frame-wise residuals $\boldsymbol{\tau}_t$. This enables tractable training of high-capacity normalizing flows while preserving the probabilistic decomposition of the model.

Autoencoder. Frames are first compressed by a convolutional autoencoder trained independently of the flow and sequence encoder. The encoder consists of an initial 7×7 convolution followed by $n_{\text{down}} = 3$ strided downsampling blocks (stride 2, 3×3 convolutions), each followed by LayerNorm and ReLU, and $n_{\text{res}} = 2$ residual blocks. All residual blocks use reflection padding and LayerNorm. The encoder output is passed through a Tanh activation to produce a continuous latent in $[-1, 1]$, which is then

floor-quantized to 256 discrete levels with a bin width of $\delta = 2/256$, using a straight-through estimator (STE) for training. The decoder mirrors the encoder with transposed convolutions for upsampling, LayerNorm, and a final Tanh output layer. With a base filter count of $n_{\text{gf}} = 16$ and $n_{\text{down}} = 3$ downsampling steps, a 64×64 input frame is compressed to a latent of shape $128 \times 8 \times 8$, giving a spatial compression factor of $8\times$ and a latent dimensionality of $D = 128 \times 8 \times 8 = 8192$. The autoencoder is trained with MSE loss using the Adam optimizer [87] at learning rate 10^{-4} until convergence, with a noise ramp that progressively introduces the dequantization noise over the first 10 epochs to match the evaluation conditions of the normalizing flow.

Sequence Encoder. The sequence encoder maps a video sequence $\boldsymbol{x}_{1:T}$ represented as AE latents (B, T, C_l, H_l, W_l) to a single global latent $\boldsymbol{s} \in \mathbb{R}^{C_l \times H_l \times W_l}$. It consists of a 1×1 convolutional channel lift followed by $n_{\text{ve}} = 1$ spatio-temporal block, each comprising a spatial residual block (two 3×3 convolutions with LayerNorm and GELU, shared across frames) and a temporal multi-head self-attention layer (4 heads) operating across the T dimension at every spatial position. A final 1×1 projection with LayerNorm maps to the output channels. The hidden dimension is 128, and no spatial downsampling is performed so the output spatial resolution matches the input. The per-frame outputs are averaged across time to produce $\boldsymbol{s}$. The encoder supports sequences of up to $T_{\max} = 25$ frames.

Conditional Normalizing Flow. The conditional normalizing flow is based on the TarFlow architecture [175], a scalable transformer-based image normalizing flow operating on AE latents. We modify the original architecture, which was conditioned on class labels, to instead condition on the image-shaped global latent $\boldsymbol{s}$ via cross-attention.

The input latent $\boldsymbol{x}_t \in \mathbb{R}^{C_l \times H_l \times W_l}$ is first patchified into a sequence of $N_p = (H_l/p)^2$ tokens of dimension $C_l \cdot p^2$ with patch size $p = 1$, giving $N_p = 64$ tokens of dimension 128 for our 8×8 latent. The flow consists of $n_{\text{blocks}} = 4$ MetaBlocks with alternating identity and flip permutations. Each MetaBlock contains $n_{\text{layers}} = 4$ AttentionBlocks. Each AttentionBlock comprises: (i) a causal masked self-attention layer with a head dimension of 64; (ii) a cross-attention layer that attends to the context tokens derived from $\boldsymbol{s}$; and (iii) a two-layer MLP with GELU activation and expansion factor 4. All attention layers use LayerNorm on inputs. Each MetaBlock implements an autoregressive affine coupling transformation

using the volume-preserving (NVP) variant, predicting both a scale and a shift for each token causally from tokens $0, \ldots, i-1$ via the attention stack. The output projection of each MetaBlock is zero-initialized, ensuring the flow starts as the identity transformation and stabilizing early training.

To condition on $\boldsymbol{s}$, the global latent is patchified with the same patch size $p = 1$ into $N_p = 64$ context tokens, projected to the model dimension via a linear layer, and augmented with learnable positional embeddings. These context tokens are used as keys and values in the cross-attention layers of each AttentionBlock, while the flow token sequence provides the queries. This allows every position in the flow to attend to all spatial positions of $\boldsymbol{s}$ at each layer.

Progressive conditioning schedule. To stabilize the conditioning on the global latent $\boldsymbol{s}$ at the beginning of training, we introduce a progressive warm-up strategy. At initialization, the flow has not yet learned to exploit $\boldsymbol{s}$ meaningfully, and conditioning directly on it can destabilize early optimization. We therefore initially condition frames on a convex combination $\rho\, \boldsymbol{s}_t + (1-\rho)\, \boldsymbol{s}$, where $\boldsymbol{s}_t$ denotes the per-frame latent from the sequence encoder and ρ follows a rapidly decaying cosine schedule from 1 to 0 over the first 30% of training. The model thus begins with frame-specific conditioning, which is easy to exploit, and progressively transitions to relying solely on the global latent $\boldsymbol{s}$ as training stabilizes.

Autoregressive Prior. For the autoregressive BPD evaluation, a temporal prior $p_\theta(\boldsymbol{\tau}_{1:T}) = p(\boldsymbol{\tau}_1) \prod_{t=2}^{T} p_\theta(\boldsymbol{\tau}_t \mid \boldsymbol{\tau}_{<t})$ is trained on top of the frozen flow residuals $\boldsymbol{\tau}_t$. The prior models $p_\theta(\boldsymbol{\tau}_t \mid \boldsymbol{\tau}_{<t})$ as $K = 4$ stacked affine coupling layers whose shift and scale parameters are predicted from a causal ConvLSTM context with hidden dimension 128. Layers use alternating fixed channel permutations (identity for the first layer and fixed random shuffles for subsequent ones) so that different channels are active across layers. The ConvLSTM operates on spatial feature maps of shape (D, H_l, W_l), preserving spatial structure so that each latent pixel receives its own shift and scale prediction. The context at step t is computed from the shifted sequence $[0, \boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_{t-1}]$ passed through the ConvLSTM in a single parallel forward pass during training, and autoregressively at test time. Optionally, the global latent $\boldsymbol{s}$ is encoded by a small CNN and added to the ConvLSTM input at every step, making the prior aware of global video structure.

Training Procedure. Training proceeds in two stages. **Stage 1** trains the autoencoder independently, as described above. **Stage 2** trains either the unconditional or conditional normalizing flow on the frozen AE latents, with a $\mathcal{N}(0, I)$ prior on the residual latents $\boldsymbol{\tau}_t$ in both cases. For the conditional model, the sequence encoder and flow are trained jointly, with the progressive cosine conditioning schedule described in Sec.5.4 applied over the first 30% of training. Both models are trained with batch size 32 and learning rate 2×10^{-3}.

The complete set of hyperparameters is summarized in Tab.C.1.

Table C.1 Hyperparameters used across all experiments.

Component	Parameter	Value
Autoencoder	Base filters n_{gf}	16
	Downsampling steps n_{down}	3
	Residual blocks n_{res}	2
	Latent shape	$128 \times 8 \times 8$
Sequence Encoder	Hidden channels	128
	Spatio-temporal blocks	1
	Attention heads	4
	Max frames T_{max}	25
Normalizing Flow	Hidden dimension	128
	MetaBlocks n_{blocks}	4
	Layers per block n_{layers}	4
	Patch size p	1
Training	Batch size	32
	Learning rate	2×10^{-3}
	Optimizer	AdamW
	β_1, β_2	0.9, 0.95
	Weight decay	10^{-4}
	Progressive schedule target	30%

Note on LLM Usage

This thesis was written in May 2026 with the assistance of a Large Language Model, specifically Claude Sonnet 4.6 by Anthropic. Large language models represent valuable tools for non-native English speakers, offering support in achieving clearer expression and appropriate academic phrasing. However, these tools also demonstrate notable limitations when it comes to generating original ideas for the kind of deep theoretical inquiry that doctoral research demands. Accordingly, LLMs were used solely to refine the presentation and language of this text. They did not contribute to its conceptual foundations, theoretical arguments, or intellectual content. All ideas, analyzes, and scholarly contributions presented in this thesis are entirely the author's own original work. More specifically, the first contribution of this thesis was written without any LLM assistance, while the two subsequent contributions were composed with the aid of LLM tools at the level of writing and language refinement only.

Bibliography

[1] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010.

[2] Siddharth Agrawal and Ambedkar Dukkipati. Deep variational inference without pixel-wise reconstruction. *arXiv preprint arXiv:1611.05209*, 2016.

[3] Niki Aifanti, Christos Papachristou, and Anastasios Delopoulos. The MUG facial expression database. In *11th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pages 1–4. IEEE, 2010.

[4] Juan F Hernandez Albarracin and Adín Ramírez Rivera. Video reenactment as inductive bias for content-motion disentanglement. *IEEE Transactions on Image Processing*, 2022.

[5] Carl Allen. Unpicking data at the seams: Understanding disentanglement in VAEs. *arXiv preprint arXiv:2410.22559*, 2024.

[6] Shun-ichi Amari, Andrzej Cichocki, and Howard Yang. A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems*, volume 8, 1995.

[7] Junwen Bai, Weiran Wang, and Carla P Gomes. Contrastively disentangled sequential variational autoencoder. In *Advances in Neural Information Processing Systems*, volume 34, pages 10105–10118, 2021.

[8] Thomas Barnett, Shruti Jain, Usha Andra, and Taru Khurana. Cisco visual networking index (VNI) complete forecast update, 2017–

2022. *Americas/EMEAR Cisco Knowledge Network (CKN) Presentation*, 1(1):2018, 2018.

[9] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[10] Nimrod Berman, Ilan Naiman, and Omri Azencot. Multifactor sequential disentanglement via structured Koopman autoencoders. *arXiv preprint arXiv:2303.17264*, 2023.

[11] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning*, volume 2, page 4, 2021.

[12] Simon Bing, Vincent Fortuin, and Gunnar Rätsch. On disentanglement in Gaussian process variational autoencoders. *arXiv preprint arXiv:2102.05507*, 2021.

[13] David E Blair. *Inversion Theory and Conformal Mapping*, volume 9. American Mathematical Society, 2000.

[14] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[15] Stella Bounareli, Christos Tzelepis, Vasileios Argyriou, Ioannis Patras, and Georgios Tzimiropoulos. One-shot neural face reenactment via finding directions in GAN's latent space. *arXiv preprint arXiv:2402.03553*, 2024.

[16] Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 38319–38331, 2022.

[17] Luitzen EJ Brouwer. Beweis der Invarianz der Dimensionenzahl. *Mathematische Annalen*, 70(2):161–165, 1911.

[18] Simon Buchholz, Michel Besserve, and Bernhard Schölkopf. Function classes for identifiable nonlinear independent component analysis. In *Advances in Neural Information Processing Systems*, volume 35, pages 16946–16961, 2022.

[19] Chris Burgess and Hyunjik Kim. 3D shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

[20] Marc-André Carbonneau, Julian Zaidi, Jonathan Boilard, and Ghyslain Gagnon. Measuring disentanglement: A review of metrics. *IEEE transactions on neural networks and learning systems*, 35(7), 2022.

[21] Hugo Caselles-Dupré, Michael Garcia Ortiz, and David Filliat. Symmetry-based disentangled representation learning requires interaction with environments. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[22] Jaehoon Cha and Jeyan Thiyagalingam. Orthogonality-enforced latent space in autoencoders: An approach to learning disentangled representations. In *Proceedings of the International Conference on Machine Learning*, pages 3913–3948. PMLR, 2023.

[23] Menglei Chai, Jian Ren, Aliaksandr Siarohin, Sergey Tulyakov, and Oliver Woodford. Motion representations for articulated animation, 2023.

[24] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[25] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *Proceedings of the IEEE International Conference on Image Processing*, pages 168–172. IEEE, 2015.

[26] Ricky TQ Chen, Jens Behrmann, David K Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for invertible generative modeling. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[27] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[28] Siyi Chen, Huijie Zhang, Minzhe Guo, Yifu Lu, Peng Wang, and Qing Qu. Exploring low-dimensional subspace in diffusion models for controllable image editing. In *Advances in Neural Information Processing Systems*, volume 37, pages 27340–27371, 2024.

[29] Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016.

[30] Pierre Comon and Christian Jutten. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, 2010.

[31] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[32] George Darmois. Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire. *Revue de l'Institut International de Statistique*, pages 2–8, 1953.

[33] Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. TCLR: Temporal contrastive learning for video representation. *Computer Vision and Image Understanding*, 219:103406, 2022.

[34] Emily L Denton et al. Unsupervised learning of disentangled representations from video. In *Advances in Neural Information Processing Systems*, 2017.

[35] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

[36] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. *arXiv preprint arXiv:1605.08803*, 2016.

[37] Michael Dorkenwald, Fanyi Xiao, Biagio Brattoli, Joseph Tighe, and Davide Modolo. SCVRL: Shuffled contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4132–4141, 2022.

[38] Emilien Dupont. Learning disentangled joint continuous and discrete representations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[39] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[40] Cian Eastwood, Julius von Kügelgen, Linus Ericsson, Diane Bouchacourt, Pascal Vincent, Bernhard Schölkopf, and Mark Ibrahim. Self-supervised disentanglement by leveraging structure in data augmentations. *arXiv preprint arXiv:2311.08815*, 2023.

[41] Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *Proceedings of the International Conference on Learning Representations*, 2018.

[42] Frederik Ebert, Chelsea Finn, Alex X Lee, and Sergey Levine. Self-supervised visual planning with temporal skip connections. In *Proceedings of the Conference on Robot Learning*, volume 12, page 23, 2017.

[43] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[44] Jana Fragemann, Lynton Ardizzone, Jan Egger, and Jens Kleesiek. Review of disentanglement approaches for medical applications – towards solving the Gordian knot of generative models in healthcare. *arXiv preprint arXiv:2203.11132*, 2022.

[45] Aviv Gabbay, Niv Cohen, and Yedid Hoshen. An image is worth more than a thousand words: Towards disentanglement in the wild. In *Advances in Neural Information Processing Systems*, volume 34, pages 9216–9228, 2021.

[46] Aviv Gabbay and Yedid Hoshen. Demystifying inter-class disentanglement. *arXiv preprint arXiv:1906.11796*, 2019.

[47] Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. Conditional neural processes. In *Proceedings of the International Conference on Machine Learning*, pages 1704–1713. PMLR, 2018.

[48] Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018.

[49] Shubhangi Ghosh, Luigi Gresele, Julius von Kügelgen, Michel Besserve, and Bernhard Schölkopf. Independent mechanism analysis and the manifold hypothesis. *arXiv preprint arXiv:2312.13438*, 2023.

[50] Gene H Golub and Charles F Van Loan. Matrix computations, 4th edition, 2013.

[51] Muhammad Waleed Gondal, Manuel Wuthrich, Djordje Miladinovic, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. In *Advances in Neural Information Processing Systems*, 2019.

[52] Luigi Gresele, Julius Von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanism analysis, a new concept? In *Advances in Neural Information Processing Systems*, volume 34, pages 28233–28248, 2021.

[53] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Fei-Fei Li, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. In *Proceedings of the European Conference on Computer Vision*, pages 393–411. Springer, 2024.

[54] Takeshi Haga, Hiroshi Kera, and Kazuhiko Kawamoto. Sequential variational autoencoder with adversarial classifier for video disentanglement. *Sensors*, 2023.

[55] Hermanni Hälvä and Aapo Hyvarinen. Hidden Markov nonlinear ICA: Unsupervised learning from nonstationary time series. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 939–948. PMLR, 2020.

[56] Jun Han, Martin Renqiang Min, Ligong Han, Li Erran Li, and Xuan Zhang. Disentangled recurrent Wasserstein autoencoder. *arXiv preprint arXiv:2101.07496*, 2021.

[57] Jan Hauke and Tomasz Kossowski. Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaestiones Geographicae*, 30(2):87–93, 2011.

[58] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022.

[59] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

[60] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the International Conference on Learning Representations* [61].

[61] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the International Conference on Learning Representations*, 2017.

[62] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[63] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[64] Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *Proceedings of the International Conference on Machine Learning*, pages 2722–2730. PMLR, 2019.

[65] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[66] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, pages 8633–8646, 2022.

[67] Daniella Horan, Eitan Richardson, and Yair Weiss. When is unsupervised disentanglement possible? In *Advances in Neural Information Processing Systems*, volume 34, pages 5150–5161, 2021.

[68] Wei-Ning Hsu and James Glass. Scalable factorized hierarchical variational autoencoder training. *arXiv preprint arXiv:1804.03201*, 2018.

[69] Wei-Ning Hsu, Yu Zhang, and James Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[70] Yaosi Hu, Zhenzhong Chen, and Chong Luo. LaMD: Latent motion diffusion for image-conditional video generation. *arXiv preprint arXiv:2304.11603*, 2023.

[71] Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron Courville. Neural autoregressive flows. In *Proceedings of the International Conference on Machine Learning*, 2018.

[72] Michael F Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics – Simulation and Computation*, 18(3):1059–1076, 1989.

[73] Sunil Hwang, Jaehong Yoon, Youngwan Lee, and Sung Ju Hwang. EVEREST: Efficient masked video autoencoder by removing redundant spatiotemporal tokens. *arXiv preprint arXiv:2211.10636*, 2022.

[74] Aapo Hyvärinen, Jarmo Hurri, and Patrik O Hoyer. Independent component analysis. In *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*, pages 151–175. Springer, 2001.

[75] Aapo Hyvärinen, Ilyes Khemakhem, and Ricardo Monti. Identifiability of latent-variable and structural-equation models: from linear to nonlinear. *Annals of the Institute of Statistical Mathematics*, 76(1):1–33, 2024.

[76] Aapo Hyvärinen, Ilyes Khemakhem, and Hiroshi Morioka. Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning. *Patterns*, 4(10), 2023.

[77] Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[78] Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ICA of temporally dependent stationary sources. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 460–469. PMLR, 2017.

[79] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

[80] Johan Ludwig William Valdemar Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30(1):175–193, 1906.

[81] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[82] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[83] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

[84] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *Proceedings of the International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018.

[85] Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. *arXiv preprint arXiv:1901.05761*, 2019.

[86] Minyoung Kim, Yuting Wang, Pritish Sahu, and Vladimir Pavlovic. Relevance factor VAE: Learning and identifying disentangled factors. *arXiv preprint arXiv:1902.01568*, 2019.

[87] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[88] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[89] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[90] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[91] Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, 2020.

[92] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 2004.

[93] Haofei Kuang, Yi Zhu, Zhi Zhang, Xinyu Li, Joseph Tighe, Sören Schwertfeger, Cyrill Stachniss, and Mu Li. Video contrastive learning with global context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3195–3204, 2021.

[94] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017.

[95] Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *Proceedings of the Conference on Causal Learning and Reasoning*, pages 428–484. PMLR, 2022.

[96] Te-Won Lee. Independent component analysis. In *Independent Component Analysis: Theory and Applications*, pages 27–66. Springer, 1998.

[97] Erich Leo Lehmann and George Casella. *Theory of Point Estimation*. Springer, 1998.

[98] José Lezama. Overcoming the disentanglement vs reconstruction trade-off via Jacobian supervision. In *Proceedings of the International Conference on Learning Representations*, 2019.

[99] Yingzhen Li and Stephan Mandt. Disentangled sequential autoencoder. *arXiv preprint arXiv:1803.02991*, 2018.

[100] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In *Proceedings of the International Conference on Machine Learning*, pages 13557–13603. PMLR, 2022.

[101] Huaize Liu, Wenzhang Sun, Qiyuan Zhang, Donglin Di, Biao Gong, Hao Li, Chen Wei, and Changqing Zou. Hi-VAE: Efficient video autoencoding with global and detailed motion. *arXiv preprint arXiv:2506.07136*, 2025.

[102] Xiao Liu, Pedro Sanchez, Spyridon Thermos, Alison Q O'Neil, and Sotirios A Tsaftaris. Learning disentangled representations in the imaging domain. *Medical Image Analysis*, 2022.

[103] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the International Conference on Machine Learning*, pages 4114–4124. PMLR, 2019.

[104] Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *Proceedings of the International Conference on Machine Learning*, pages 6348–6359. PMLR, 2020.

[105] Francesco Locatello, Michael Tschannen, Stefan Bauer, Gunnar Rätsch, Bernhard Schölkopf, and Olivier Bachem. Disentangling factors of variation using few labels. *arXiv preprint arXiv:1905.01258*, 2019.

[106] Yin-Jyun Luo, Sebastian Ewert, and Simon Dixon. Towards robust unsupervised disentanglement of sequential data – a case study using music audio. *arXiv preprint arXiv:2205.05871*, 2022.

[107] Xuezhe Ma, Xiang Kong, Shanghang Zhang, and Eduard Hovy. Decoupling global and local representations via invertible generative flows. *arXiv preprint arXiv:2004.11820*, 2020.

[108] Joseph Marino, Lei Chen, Jiawei He, and Stephan Mandt. Improving sequential latent variable models with autoregressive flows. In

Symposium on Advances in Approximate Bayesian Inference, pages 1–16. PMLR, 2020.

[109] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dSprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

[110] Graziano Mita, Maurizio Filippone, and Pietro Michiardi. An identifiable double VAE for disentangled representations. In *Proceedings of the International Conference on Machine Learning*, pages 7769–7779. PMLR, 2021.

[111] Rogan Morrow and Wei-Chen Chiu. Variational autoencoders with normalizing flow decoders. *arXiv preprint arXiv:2004.05617*, 2020.

[112] Ilan Naiman, Nimrod Berman, and Omri Azencot. Sample and predict your latent: Modality-free sequential disentanglement via contrastive estimation. *arXiv preprint arXiv:2305.15924*, 2023.

[113] Erkki Oja, Juha Karhunen, Harri Valpola, Jaakko Särelä, Mika Inki, Antti Honkela, Alexander Ilin, Karthikesh Raju, Tapani Ristaniemi, and Ella Bingham. Independent component analysis and blind source separation. *Helsinki University of Technology, Espoo, Finland, Technical Report*, 2003.

[114] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[115] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

[116] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[117] Jungin Park, Jiyoung Lee, Ig-Jae Kim, and Kwanghoon Sohn. Probabilistic representations for video contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14711–14721, 2022.

[118] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* [119].

[119] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[120] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[121] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[122] William Peebles, John Peebles, Jun-Yan Zhu, Alexei Efros, and Antonio Torralba. The Hessian penalty: A weak prior for unsupervised disentanglement. In *Proceedings of the European Conference on Computer Vision*, pages 581–597. Springer, 2020.

[123] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

[124] Ryan Po, Guandao Yang, Kfir Aberman, and Gordon Wetzstein. Orthogonal adaptation for modular customization of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7964–7973, 2024.

[125] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021.

[126] Robin Quessard, Thomas Barrett, and William Clements. Learning disentangled representations and group structure of dynamical en-

vironments. In *Advances in Neural Information Processing Systems*, volume 33, pages 19727–19737, 2020.

[127] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[128] Scott E Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[129] Patrik Reizinger, Luigi Gresele, Jack Brady, Julius Von Kügelgen, Dominik Zietlow, Bernhard Schölkopf, Georg Martius, Wieland Brendel, and Michel Besserve. Embrace the gap: VAEs perform independent mechanism analysis. In *Advances in Neural Information Processing Systems*, volume 35, pages 12040–12057, 2022.

[130] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the International Conference on Machine Learning*, pages 1530–1538. PMLR, 2015.

[131] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[132] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[133] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[134] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[135] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local SVM approach. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 3, pages 32–36. IEEE, 2004.

[136] Cuifeng Shen, Lumin Xu, Xingguo Zhu, and Gengdai Liu. Autoregressive video autoencoder with decoupled temporal and spatial context. *arXiv preprint arXiv:2512.11293*, 2025.

[137] Jonathon Shlens. A tutorial on independent component analysis. *arXiv preprint arXiv:1404.2986*, 2014.

[138] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[139] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Advances in Neural Information Processing Systems*, 2019.

[140] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[141] Mathieu Cyrille Simon, Pascal Frossard, and Christophe De Vleeschouwer. Sequential representation learning via static-dynamic conditional disentanglement. In *Proceedings of the European Conference on Computer Vision*, pages 110–126. Springer, 2024.

[142] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using LSTMs. In *Proceedings of the International Conference on Machine Learning*, pages 843–852. PMLR, 2015.

[143] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. *arXiv preprint arXiv:2104.15069*, 2021.

[144] Sana Tonekaboni, Chun-Liang Li, Sercan O Arik, Anna Goldenberg, and Tomas Pfister. Decoupling local and global representations of time series. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2022.

[145] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised

video pre-training. In *Advances in Neural Information Processing Systems*, volume 35, pages 10078–10093, 2022.

[146] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MoCoGAN: Decomposing motion and content for video generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1526–1535, 2018.

[147] Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. In *Advances in Neural Information Processing Systems*, 2020.

[148] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008.

[149] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. *arXiv preprint arXiv:1706.08033*, 2017.

[150] Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. In *Advances in Neural Information Processing Systems*, volume 36, pages 48603–48638, 2023.

[151] Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, 2021.

[152] Elif Vural and Pascal Frossard. Learning pattern transformation manifolds for classification. In *Proceedings of the IEEE International Conference on Image Processing*. IEEE, 2012.

[153] Joseph L Walsh. History of the Riemann mapping theorem. *The American Mathematical Monthly*, 80(3):270–276, 1973.

[154] Tan Wang, Zhongqi Yue, Jianqiang Huang, Qianru Sun, and Hanwang Zhang. Self-supervised learning disentangled group representation as feature. In *Advances in Neural Information Processing Systems*, volume 34, pages 18225–18240, 2021.

[155] Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. Disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9677–9696, 2024.

[156] Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. Disentangled representation learning. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* [155], pages 9677–9696.

[157] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. G3AN: Disentangling appearance and motion for video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[158] Yuchi Wang, Junliang Guo, Xinyi Xie, Tianyu He, Xu Sun, and Jiang Bian. VidTwin: Video VAE with decoupled structure and dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22922–22932, 2025.

[159] Yuxiang Wei, Yupeng Shi, Xiao Liu, Zhilong Ji, Yuan Gao, Zhongqin Wu, and Wangmeng Zuo. Orthogonal Jacobian regularization for unsupervised disentanglement in image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6721–6730, 2021.

[160] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. *arXiv preprint arXiv:1906.02634*, 2019.

[161] Christina Winkler, Daniel Worrall, Emiel Hoogeboom, and Max Welling. Learning likelihoods with conditional normalizing flows. *arXiv preprint arXiv:1912.00042*, 2019.

[162] Quanhan Xi, Sebastian Gonzalez, and Benjamin Bloem-Reddy. Triangular monotonic generative models can perform causal discovery. In *Causal Representation Learning Workshop at NeurIPS 2023*, 2023.

[163] Jianwen Xie, Ruiqi Gao, Zilong Zheng, Song-Chun Zhu, and Ying Nian Wu. Motion-based generator model: Unsupervised disentanglement of appearance, trackable and intrackable motions in dynamic patterns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[164] Yazhou Xing, Yang Fei, Yingqing He, Jingye Chen, Jiaxin Xie, Xiaowei Chi, and Qifeng Chen. VideoVAE+: Large motion video

autoencoding with cross-modal video VAE. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17951–17960, 2025.

[165] Danru Xu, Dingling Yao, Sébastien Lachapelle, Perouz Taslakian, Julius Von Kügelgen, Francesco Locatello, and Sara Magliacane. A sparsity principle for partially observable causal representation learning. *arXiv preprint arXiv:2403.08335*, 2024.

[166] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. VideoGPT: Video generation using VQ-VAE and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

[167] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. CausalVAE: Structured causal disentanglement in variational autoencoder. *arXiv preprint arXiv:2004.08697*, 2020.

[168] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. CausalVAE: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[169] Tao Yang, Xuanchi Ren, Yuwang Wang, Wenjun Zeng, and Nanning Zheng. Towards building a group-based unsupervised representation disentanglement framework. *arXiv preprint arXiv:2102.10303*, 2021.

[170] Zhuoqian Yang, Wentao Zhu, Wayne Wu, Chen Qian, Qiang Zhou, Bolei Zhou, and Chen Change Loy. TransMoMo: Invariance-driven unsupervised video motion retargeting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[171] Xi Ye and Guillaume-Alexandre Bilodeau. A unified model for continuous conditional video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[172] Xi Ye and Guillaume-Alexandre Bilodeau. Continuous conditional video synthesis by neural processes. *Computer Vision and Image Understanding*, 259:104387, 2025.

[173] Dacheng Yin, Xuanchi Ren, Chong Luo, Yuwang Wang, Zhiwei Xiong, and Wenjun Zeng. Retriever: Learning content-style

representation as a token-level bipartite graph. *arXiv preprint arXiv:2202.12307*, 2022.

[174] Sihyun Yu, Weili Nie, De-An Huang, Boyi Li, Jinwoo Shin, and Anima Anandkumar. Efficient video diffusion models via content-frame motion-latent decomposition. *arXiv preprint arXiv:2403.14148*, 2024.

[175] Shuangfei Zhai, Ruixiang Zhang, Preetum Nakkiran, David Berthelot, Jiatao Gu, Huangjie Zheng, Tianrong Chen, Miguel Angel Bautista, Navdeep Jaitly, and Josh Susskind. Normalizing flows are capable generative models. *arXiv preprint arXiv:2412.06329*, 2024.

[176] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Towards deeper understanding of variational autoencoding models. *arXiv preprint arXiv:1702.08658*, 2017.

[177] Sijie Zhao, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Muyao Niu, Xiaoyu Li, Wenbo Hu, and Ying Shan. CV-VAE: A compatible video VAE for latent generative video models. In *Advances in Neural Information Processing Systems*, volume 37, pages 12847–12871, 2024.

[178] Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ICA: Sparsity and beyond. In *Advances in Neural Information Processing Systems*, volume 35, pages 16411–16422, 2022.

[179] Xinqi Zhu, Chang Xu, and Dacheng Tao. Commutative Lie group VAE for disentanglement learning. In *Proceedings of the International Conference on Machine Learning*, pages 12924–12934. PMLR, 2021.

[180] Yizhe Zhu, Martin Renqiang Min, Asim Kadav, and Hans Peter Graf. S3VAE: Self-supervised sequential VAE for representation disentanglement and data generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.