Vocabulary-free few-shot learning for Vision-Language Models

Maxime Zanella*[1,2] Clément Fuchs*[1] Ismail Ben Ayed[3] Christophe De Vleeschouwer[1]
[1]UCLouvain, Belgium [2]UMons, Belgium [3]ÉTS Montreal, Canada

Abstract

Recent advances in few-shot adaptation for Vision-Language Models (VLMs) have greatly expanded their ability to generalize across tasks using only a few labeled examples. However, existing approaches primarily build upon the strong zero-shot priors of these models by leveraging carefully designed, task-specific prompts. This dependence on predefined class names can restrict their applicability, especially in scenarios where exact class names are unavailable or difficult to specify. To address this limitation, we introduce vocabulary-free few-shot learning for VLMs, a setting where target class instances - that is, images - are available but their corresponding names are not. We propose Similarity Mapping (SiM), a simple yet effective baseline that classifies target instances solely based on similarity scores with a set of generic prompts (textual or visual), eliminating the need for carefully handcrafted prompts. Although conceptually straightforward, SiM demonstrates strong performance, operates with high computational efficiency (learning the mapping typically takes less than one second), and provides interpretability by linking target classes to generic prompts. We believe that our approach could serve as an important baseline for future research in vocabulary-free few-shot learning. Code available at https://github.com/MaxZanella/vocabulary-free-FSL.

1. Introduction

Vision-Language Models (VLMs), such as CLIP [26], have become essential for cross-modal learning. Such models align images and text through large-scale contrastive training. One of their key strengths is zero-shot classification, which enables them to categorize images solely based on textual prompts that describe target classes. This ability has led to impressive results and motivated recent developments of few-shot learning techniques, which further adapt VLMs to new tasks using only a small number of labeled examples. To achieve this, current few-shot methods build on this strong zero-shot classification, either by incorporating class name tokens into learnable prompts [17, 40, 41], by employing handcrafted prompts in combination with adapters [36, 38], or by using low-rank fine-tuning [37], to name a few approaches.

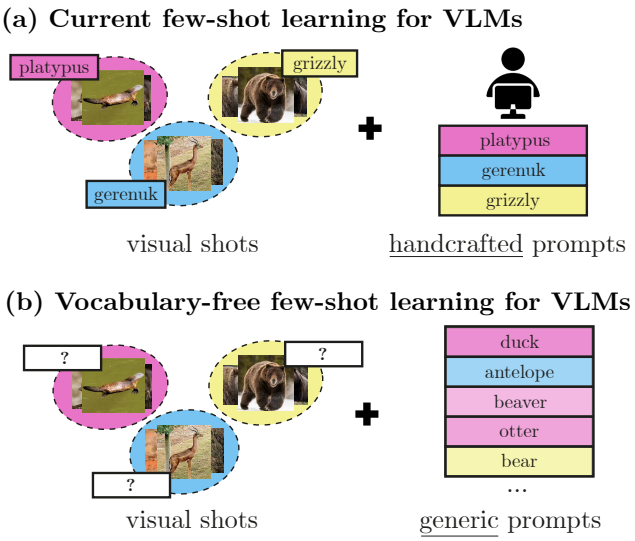


Figure 1. Current few-shot learning methods assume that target class names are known, often requiring manually fine-tuned prompts. In *vocabulary-free few-shot learning*, we remove this constraint and rely solely on generic prompts (e.g., derived from ImageNet classes).

However, in many cases, class definitions may be vague or ambiguous, making it difficult to design meaningful prompts. Secondly, some classes may require long and complex descriptions that a single prompt may not fully capture, making it preferable to break them down into smaller and more interpretable components. Thirdly, new concepts may emerge, although unknown during VLM pretraining. These practical challenges can hinder direct zero-shot classification and, consequently, limit existing few-shot learning methods that rely on predefined class names. To overcome these limitations, we introduce *vocabulary-free few-shot learning*, a framework in which visual instances of the target classes are available, but their corresponding names are not. A comparison with standard few-

* Equal contributions and corresponding authors. {maxime.zanella,clement.fuchs}@uclouvain.be

shot learning for VLMs is provided in Figure 1.

Without predefined class names, current few-shot methods for VLMs become inapplicable, requiring alternative ways to discriminate classes. In our work, we propose the use of generic textual prompts derived from the ImageNet class names (Table 1), while illustrating how broader concepts, such as those from the Wordnet lexical database [23], could also be explored. Beyond classification, our approach enhances interpretability by linking target classes to these generic prompts representing meaningful concepts (see Figure 3). This, in turn, may provide a semantic understanding of the target classes. For example, in our experiments, the target class gerenuk is linked to impala and gazelle, other antelope species (Figure 3a).

To take advantage of the generic textual prompts, we propose to learn a linear mapping between those prompts and the target classes. Given a small number of labeled examples (shots) per target class, we estimate a mapping that projects similarity scores between images and generic textual prompts onto class assignments. Our approach shares similarities with label mapping techniques used in visual reprogramming [5, 12, 33], where a mapping function aligns pre-trained model outputs with new task labels. Our method is highly efficient, operates as a black-box model (relying solely on similarity scores rather than direct access to textual or visual embeddings), and requires minimal computational overhead.

Beyond its current formulation, we believe that *vocabulary-free few-shot learning* and our proposed baseline open up several promising research directions. They include the adaptation of recent advances in few-shot learning to the vocabulary-free problem, the expansion of the set of generic prompts, the integration of richer textual and image-based databases or accounting for prior knowledge about class relationships. Finally, assigning meaningful names to the target classes remains an open problem.

Contributions. In this work, we introduce *vocabulary-free few-shot learning*, a new paradigm where target class visual instances are available, but their exact names are not—challenging the conventional reliance on predefined class names in few-shot Vision-Language Model adaptation. To address this problem, we propose a simple yet effective baseline that learns a linear mapping between generic prompts and target classes, enabling classification without explicit textual target labels. In addition, our approach provides a desirable interpretability property, as the learned mapping offers insights into how target classes relate to known concepts. Finally, we outline several research directions for future improvements, including refining the few-shot learning algorithm. expanding the diversity of generic prompts (texts and/or images), and enabling the fine-grained naming of the target classes.

2. Related works

Few-shot learning in Vision-Language. Adapting Vision-Language Models (VLMs) to new tasks with minimal supervision has become a predominant area of interest. While their zero-shot capabilities enable broad generalization, fine-tuning on a few labeled examples can still significantly improve performance. Prompt tuning has emerged as a dominant strategy, refining textual and/or visual embeddings to enhance adaptation [3, 6, 11, 17, 18, 21, 35, 39–41]. Methods like CoOp [40] optimize learnable common continuous tokens attached to the class names, described as a context optimization. MaPLe [17] extends this strategy by introducing learnable visual tokens in addition to textual ones. A second line of research focuses on adapter-based methods, which modify a small subset of parameters to improve efficiency [14, 36, 38]. Tip-Adapter [38] leverages memory-based caching, combining stored feature representations with the original zero-shot prediction. TaskRes [36] introduces task-specific residual tuning, adapting the initial text embedding of each class prompt. A third path of investigation considers low-rank fine-tuning, exemplified by CLIP-LoRA [37], which applies low-rank adaptation within both text and visual encoders. All these methods inherently rely on predefined class names, making them unsuitable when explicit class names are unavailable.

Vocabulary-free classification. Most existing methods for VLMs assume that class names are available at test time, as they are essential for generating textual prompts. However, this assumption becomes impractical when the semantic context is unknown or evolving. Recent works [8, 9] have tackled this issue by assigning names to images from an unconstrained set of semantic concepts. A related approach is retrieval augmented models, based on large-scale image-text pairs database such as LAION-5B [27], Yfcc100m [32], Conceptual Captions [28], or the Public Multimodal Datasets (PMD) [29]. These datasets can be used to retrieve similar images and their corresponding captions [20]. Our proposed *vocabulary-free few-shot learning* method named SiM differs from these approaches as it focuses on learning a classifier from groups of images (the visual shots) rather than improving single-image classification.

Label mapping Label mapping (LM) has emerged as a core component of visual reprogramming strategies [5, 12, 33], which focus on reconfiguring a pretrained model for arbitrary downstream tasks using a trainable transformation of the input images, together with an LM function. The latter is needed because the label spaces are often distinct between the pre-training and downstream tasks, which ne-

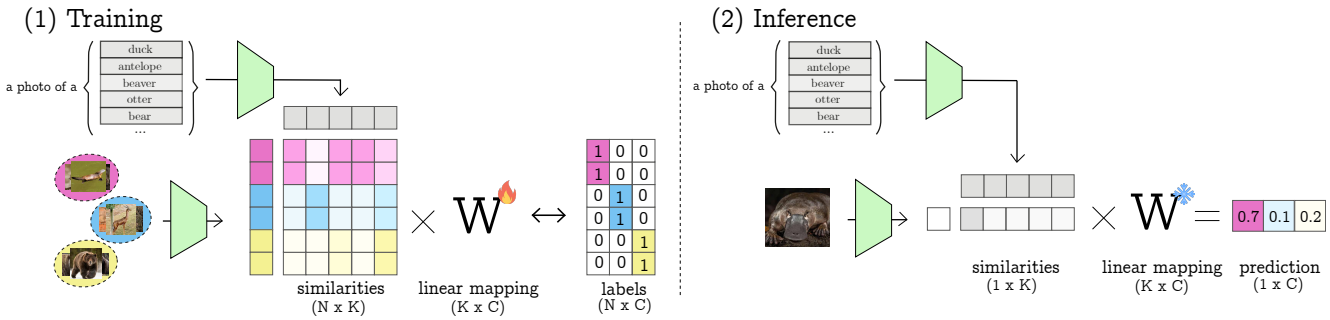


Figure 2. Summary of our approach with three target classes. (1) At training time, a linear operator W is learned via least-squares minimization to map the similarities of K generic prompts and the N visual shots to their corresponding class index (i.e., one-hot encoding). (2) At inference time, the operator W can be used to get a prediction on each test image.

cessitates to map outputs of the pretrained model to downstream labels. Initially, [12] adopted a random mapping strategy, and fully relied on the trainable transformation for accuracy. Later, [33] introduced frequency label mapping (FLM), which matches the labels of the pre-trained task to each downstream label based on the model's prediction frequencies. In the context of visual reprogramming, [5] recomputed the LM function as the training of the input transformation progressed, yielding the iterative label mapping algorithm. Recently, [4] introduced bayesian-guided label mapping (BLM), which provides a bayesian framework for estimating the LM function, leveraging the joint distribution of the pretrained label predictions and ground-truth target classes. As these methods share similarities with our approach in learning a mapping between the pretrained model outputs and target classes, we report their performances in our main experimental results for comparison.

3. Methodology

3.1. Preliminaries

To fully understand our approach for *vocabulary-free few-shot learning* for vision-language models (VLMs), we start by defining the key elements of the usual vision-language classification framework. At its core, CLIP embeds both images and textual descriptions into a common latent space, enabling alignment and measurement of their similarities. Given a set of predefined C classes with known names, the framework relies on generating so-called textual prompts, such as "a photo of a $\{\text{class}_c\}$". These textual descriptions are first tokenized and then processed by the textual encoder into normalized embeddings $\mathbf{t}_c$ on the unit hypersphere $\mathcal{S}_{d-1} = \{u \in \mathbb{R}^d; ||u||_2 = 1\}$. Similarly, the image $\mathbf{x}_i$ is transformed by the visual encoder to produce the representation $\mathbf{f}_i \in \mathcal{S}_{d-1}$. This allows for the computation of similarity scores

$$l_{i,c} = \mathbf{f}_i^\top \mathbf{t}_c. \tag{1}$$

Finally, the predicted class for image x_i is obtained by selecting the highest similarity score, i.e. $\hat{c} = \text{argmax}_c\ l_{i,c}$. Note this approach is not feasible in the vocabulary-free setting since class names $\{\text{class}_c\}$ are unavailable, preventing the computation of embeddings $\boldsymbol{t}_c$.

3.2. Our approach

Our approach, which we dub Similarities Mapping (SiM), learns a matrix W to linearly map similarity scores computed with a predefined set of generic prompts to a one-hot representation of the classes, defined by a few-shot set of images. The whole pipeline is summarized in Figure 2.

Generic prompts. Given an image x_i, since the embeddings $\boldsymbol{t}_c$ are unavailable, similarities scores $l_{i,c}$ (Eq. (1)) cannot be computed. Rather, the similarity scores are computed with respect to a set of K predefined embeddings $\boldsymbol{t}_k \in \mathcal{S}_{d-1}$, with $1 \leq k \leq K$. $\boldsymbol{t}_k$ can be obtained from a set of arbitrary and generic textual prompts or from other images. The task is then to obtain a class prediction from the scores $l_{i,k}$, given the knowledge of similarity scores for a few-shot set of images $S = \{x_j; 1 \leq j \leq N\}$. Formally, these similarity scores can be grouped into a matrix

$$L = \begin{pmatrix} \mathbf{f}_1^T\mathbf{t}_1 & \dots & \mathbf{f}_1^T\mathbf{t}_K \\ \dots & \dots & \dots \\ \mathbf{f}_N^T\mathbf{t}_1 & \dots & \mathbf{f}_N^T\mathbf{t}_K \end{pmatrix} \in \mathbb{R}^{N\times K}. \tag{2}$$

Learning from the few-shot set. We can then estimate $W \in \mathbb{R}^{K\times C}$ such that $Y \approx LW$, where $Y \in \mathbb{R}^{N\times C}$ is the matrix whose lines correspond to one-hot encoded labels of the few-shot images. This is achieved by minimizing a least-squares objective with Tikhonov regularization [15], i.e.,

$$W = \arg\min_W ||Y - LW||_F^2 + \lambda||W||_F^2, \tag{3}$$

λ being an hyperparameter. Setting the gradient of the objective in Eq. (3) to 0 yields

$$W = (L^T L + \lambda I_K)^{-1} Y. \tag{4}$$

Classification of test images. For a test image x_i, we get the similarities $l_{i,k} = \mathbf{f}_i^T \mathbf{t}_k$ with respect to the predefined set of arbitrary and generic prompts $\{\mathbf{t}_k\}_{1 \leq k \leq K}$. We then use W to map these similarities to the target class scores

$$s_{i,c} = \sum_{k=1}^{K} w_{k,c} l_{i,k}. \tag{5}$$

The predicted class is then $\hat{c} = \arg\max_c s_{i,c}$. By investigating the relative importance of the weights $w_{k,c}$ for a given target class c, we can find the embeddings most aligned in the set of K arbitrary and general prompts, potentially providing a high-level semantic understanding of the target class (see Figure 3).

Interpretation of Eq. (3) as an unsupervised clustering objective. The problem in Eq. (3) could be viewed as an unsupervised clustering objective akin to K-means [1], but operating on the feature vectors of the visual shots, $(\mathbf{f}_j^T \mathbf{t}_k)_{1 \leq k \leq K}$, $1 \leq j \leq N$, and using a fixed point-to-cluster assignment variables Y. The matrix factorization formulation of K-means in [1] enables to see this perspective. Indeed, the column vectors of matrix W, which are the optimization variables, could be viewed as cluster prototypes. So, essentially, our method could be viewed as an unsupervised clustering of the visual shots based on features driven from their similarity with the generic set of prompts.

4. Experimental validation

4.1. Comparative methods

One-to-One mapping. This approach seeks to associate a single generic prompt to each target class c, using similarities matrix L (Eq. (2)). Specifically, to learn the mapping, we adopt the *Frequency Label Mapping* (FLM) algorithm presented in [33].

Bayesian label mapping. Bayesian label mapping (BLM) was introduced in [4] and leverages the joint distribution of labels to learn a more flexible many-to-many mapping.

Centroids. To further validate our approach, we compare it to a vision-only few-shot learning baseline inspired by [30]. This method follows the usual classification framework of VLMs (see Eq. (1)) but replaces the textual prompts $\boldsymbol{t}_c$ with the sample mean of visual features from images belonging to class c. Note that this approach does not make use of the generic prompts and does not provide semantic insights into target classes.

Standard few-shot learning for VLMs. Additionally we provide results for few-shot methods requiring the class names. Because of the large popularity of these fields, we report some of the most popular methods i.e., the text prompt tuning method CoOp [40], the vision and text prompt tuning MaPLe [17], the cache-based method Tip-Adapter [38], the adapter-based method TaskRes [36] and the low-rank adaptation method CLIP-LoRA [37]. These methods serve as upper-bound references in our experiments, as they benefit from access to predefined class names and can therefore leverage zero-shot prediction, unlike vocabulary-free approaches.

4.2. Experimental setting

Datasets. We adapt the setting of previous works [40] and use 11 datasets for image classification tasks, namely ImageNet [10] a large-scale benchmark for object recognition, SUN397 [34] for fine-grained classification of scenes, Aircraft [22] for classification of aircraft types, EuroSAT [16] for satellite imagery, StanfordCars [19] for car models, Food101 [2] for food items, Pets [25] for pet types, Flower102 [24] for flower species, Caltech101 [13] for a variety of general objects, DTD [7] for texture types and UCF101 [31] for action recognition.

ImageNet class names for prompts. For the main results, we use the classes of ImageNet [10] to generate $K = 1000$ prompts of the form "a photo of a $\{\text{class}_k\}$.", and subsequently obtain the associated embeddings $\boldsymbol{t}_k$.

ImageNet images for prompts. To demonstrate that similarities can be computed using a wide variety of prompts, we investigate an alternative setting where image-based representations replace textual prompts. As discussed later, this highlights that improving the choice of prompts could be an interesting direction for future research. In this setting, for each random seed we randomly draw a single image for each of the $K = 1000$ classes, which are then processed by the visual encoder to yield the normalized embeddings $\boldsymbol{t}_k$. Hence, the similarities in the matrix L are images to images rather than images to text.

Wordnet vocabulary for prompts. In this setting, we select all words in Wordnet [23] which are related to at least one of the words in ["building", "vehicle", "food", "flower", "animal", "texture", "action", "furniture"], resulting in a vocabulary of $K = 16452$ words. We then generate prompts of the form "a photo of a $\{\text{word}_k\}$." which are processed by the text encoder and normalized to yield the embeddings $\boldsymbol{t}_k$.

Table 1. Detailed results for the 10 datasets for two CLIP backbones. Top-1 accuracy averaged over 3 random seeds is reported. Highest value for Vocabulary-free methods is highlighted in **bold**.

(a) Results for the CLIP ViT-B/16 backbone.

	Method	Vocabulary-free	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Average
	CLIP	✗	62.6	24.7	47.5	65.3	86.1	89.1	71.4	92.9	43.6	66.7	65.0
4-shot	CoOp	✗	69.7	30.9	69.7	74.4	84.5	92.5	92.2	94.5	59.5	77.6	74.6
	TIP-Adapter-F	✗	70.8	35.7	76.8	74.1	86.5	91.9	92.1	94.8	59.8	78.1	76.1
	TaskRes	✗	72.7	33.4	74.2	76.0	86.0	91.9	85.0	95.0	60.1	76.2	75.1
	MaPLe	✗	71.4	30.1	69.9	70.1	86.7	93.3	84.9	95.0	59.0	77.1	73.8
	CLIP-LoRA	✗	72.8	37.9	84.9	77.4	82.7	91.0	93.7	95.2	63.8	81.1	78.1
	One-to-One	✓	20.5	2.3	27.9	4.3	16.9	68.2	11.2	76.3	23.8	39.7	29.1
	BLM	✓	26.4	1.6	38.1	2.7	19.0	63.4	17.5	81.9	34.1	45.1	33.0
	centroids	✓	60.9	32.4	70.1	57.8	71.0	64.6	**91.2**	91.3	53.9	70.9	66.4
	SiM (ours)	✓	**62.7**	**33.2**	**75.8**	**60.5**	**75.4**	**79.2**	89.7	**93.2**	**59.0**	**73.8**	**70.2**
8-shot	CoOp	✗	71.9	38.5	77.1	79.0	82.7	91.3	94.9	94.5	64.8	80.0	77.5
	TIP-Adapter-F	✗	73.5	39.5	81.3	78.3	86.9	91.8	94.3	95.2	66.7	82.0	79.0
	TaskRes	✗	74.6	40.3	77.5	79.6	86.4	92.0	96.0	95.3	66.7	81.6	79.0
	MaPLe	✗	73.2	33.8	82.8	71.3	87.2	93.1	90.5	95.1	63.0	79.5	77.0
	CLIP-LoRA	✗	74.7	45.7	89.7	82.1	83.1	91.7	96.3	95.6	67.5	84.1	81.1
	One-to-One	✓	22.6	2.4	30.2	4.5	18.5	69.4	12.8	77.2	28.5	41.3	30.7
	BLM	✓	28.4	1.9	41.2	2.6	21.4	59.4	19.1	86.2	39.7	45.1	34.5
	centroids	✓	66.8	36.9	74.3	65.7	77.6	74.6	**94.1**	92.7	60.4	76.2	71.9
	SiM (ours)	✓	**67.2**	**38.4**	**80.1**	**68.9**	**80.1**	**84.6**	92.8	**94.8**	**64.5**	**77.0**	**74.9**
16-shot	CoOp	✗	74.9	43.2	85.0	82.9	84.2	92.0	96.8	95.8	69.7	83.1	80.8
	TIP-Adapter-F	✗	76.0	44.6	85.9	82.3	86.8	92.6	96.2	95.7	70.8	83.9	81.5
	TaskRes	✗	76.1	44.9	82.7	83.5	86.9	92.4	97.5	95.8	71.5	84.0	81.5
	MaPLe	✗	74.5	36.8	87.5	74.3	87.4	93.2	94.2	95.4	68.4	81.4	79.3
	CLIP-LoRA	✗	76.1	54.7	92.1	86.3	84.2	92.4	98.0	96.4	72.0	86.7	83.9
	One-to-One	✓	24.0	2.4	34.2	4.9	18.9	71.9	14.0	78.0	30.9	43.4	32.3
	BLM	✓	29.3	1.1	36.0	2.5	21.3	49.6	20.1	86.1	44.3	47.8	33.8
	centroids	✓	**70.1**	40.7	76.6	71.6	80.9	78.8	**95.5**	93.8	63.1	77.4	74.8
	SiM (ours)	✓	**70.1**	**43.7**	**85.6**	**74.8**	**82.8**	**88.1**	**95.5**	**95.3**	**69.9**	**79.0**	**78.5**

(b) Results for the CLIP ViT-L/14 backbone.

	Method	Vocabulary-free	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Average
	CLIP	✗	67.6	32.6	58.0	76.8	91.0	93.6	79.4	94.9	53.6	74.2	72.2
4-shot	CoOp	✗	73.7	41.8	77.9	82.6	88.8	94.7	94.9	96.1	64.3	83.6	79.8
	TIP-Adapter-F	✗	74.1	47.4	81.4	82.3	91.2	94.0	95.5	96.5	64.4	83.9	81.1
	TaskRes	✗	74.9	42.5	76.6	83.6	90.7	94.4	90.3	96.5	65.4	80.1	79.5
	MaPLe	✗	76.0	40.4	74.6	80.3	91.5	95.0	93.2	97.0	64.5	82.8	79.5
	CLIP-LoRA	✗	76.7	48.9	86.4	85.2	89.6	93.9	97.4	97.2	70.4	86.4	83.2
	One-to-One	✓	23.7	2.6	37.2	5.1	17.7	71.3	15.9	73.8	22.5	40.4	31.0
	BLM	✓	29.8	2.6	45.6	4.2	21.3	68.6	21.6	86.3	33.4	49.7	36.3
	centroids	✓	66.1	42.5	76.0	69.4	82.0	77.7	**96.3**	94.4	58.4	78.7	74.2
	SiM (ours)	✓	**67.2**	**43.0**	**80.9**	**72.5**	**85.6**	**87.3**	**96.3**	**95.8**	**61.4**	**81.8**	**77.2**
8-shot	CoOp	✗	75.5	48.7	81.4	85.9	89.2	94.5	97.5	96.5	68.8	86.0	82.4
	TIP-Adapter-F	✗	76.7	50.4	84.9	85.9	91.4	94.1	97.3	96.9	71.2	86.2	83.5
	TaskRes	✗	76.0	51.1	81.1	85.7	91.1	94.5	96.7	96.9	69.4	85.6	82.8
	MaPLe	✗	77.2	42.9	80.7	81.8	90.1	95.0	95.8	96.8	69.5	85.1	81.5
	CLIP-LoRA	✗	78.0	57.5	90.0	88.7	89.7	94.2	98.0	97.0	72.2	88.3	85.4
	One-to-One	✓	25.5	2.8	42.4	5.2	19.9	70.6	16.3	76.8	26.1	43.7	32.9
	BLM	✓	32.1	1.5	52.8	3.2	23.3	63.4	25.4	89.2	39.3	50.4	38.0
	centroids	✓	**71.9**	47.0	79.8	76.8	86.3	85.6	97.4	95.6	65.3	82.4	78.8
	SiM (ours)	✓	71.6	**47.6**	**85.2**	**79.8**	**88.4**	**91.6**	**97.5**	**96.9**	**68.3**	**83.3**	**81.0**
16-shot	CoOp	✗	77.9	53.0	86.7	87.4	90.2	94.5	98.6	97.5	73.7	86.7	84.6
	TIP-Adapter-F	✗	79.6	55.8	86.1	88.1	91.6	94.6	98.3	97.5	74.0	87.4	85.3
	TaskRes	✗	76.9	55.0	84.3	87.6	91.5	94.7	97.8	97.3	74.4	86.6	84.6
	MaPLe	✗	78.8	46.3	85.4	83.6	92.0	95.4	97.4	97.2	72.7	86.5	83.5
	CLIP-LoRA	✗	79.4	66.2	93.1	90.9	89.9	94.3	99.0	97.3	76.5	89.9	87.7
	One-to-One	✓	26.7	2.9	47.1	5.9	20.5	72.8	18.6	77.1	29.4	43.6	34.4
	BLM	✓	34.5	3.1	53.4	2.7	25.0	58.8	24.5	89.9	43.7	50.8	38.7
	centroids	✓	**74.9**	51.8	81.8	81.3	88.2	88.6	98.5	96.6	67.3	83.8	81.3
	SiM (ours)	✓	74.3	**52.8**	**90.1**	**84.0**	**89.6**	**93.5**	**98.8**	**97.3**	**73.1**	**85.7**	**83.9**

Table 2. Detailed results for the 10 datasets for two CLIP backbones. Top-1 accuracy averaged over 3 random seeds is reported. Embeddings $\boldsymbol{t}_k$ are obtained from imagenet classes, imagenet images (images) or a subset of 16452 words from Wordnet (Wordnet).

(a) Results for the CLIP ViT-B/16 backbone.

	Method	Vocabulary-free	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Average
	CLIP	✗	62.6	24.7	47.5	65.3	86.1	89.1	71.4	92.9	43.6	66.7	65.0
4-shot	SiM (images)	✓	61.9	29.7	71.5	56.4	73.6	66.1	87.3	92.1	55.7	71.2	66.6
	SiM (Wordnet)	✓	62.2	30.7	74.2	**60.8**	**75.4**	**79.8**	**89.9**	92.8	58.9	72.2	69.7
	SiM	✓	**62.7**	**33.2**	**75.8**	60.5	**75.4**	79.2	89.7	**93.2**	**59.0**	**73.8**	**70.2**
8-shot	SiM (images)	✓	67.1	37.5	70.2	65.4	79.2	77.8	92.4	94.7	61.9	75.6	73.1
	SiM (Wordnet)	✓	66.9	37.4	**81.0**	**69.1**	**80.4**	**84.8**	**93.6**	94.4	**66.0**	76.3	**75.0**
	SiM	✓	**67.2**	**38.4**	80.1	68.9	80.1	84.6	92.8	**94.8**	64.5	**77.0**	74.9
16-shot	SiM (images)	✓	70.0	**44.2**	**85.7**	73.4	82.2	83.9	94.8	95.3	69.6	**79.6**	77.9
	SiM (Wordnet)	✓	**70.3**	44.0	**85.7**	74.7	**82.8**	**88.1**	**95.5**	**95.7**	**70.8**	79.1	**78.7**
	SiM	✓	70.1	43.7	85.6	**74.8**	**82.8**	**88.1**	**95.5**	95.3	69.9	79.0	78.5

(b) Results for the CLIP ViT-L/14 backbone.

	Method	Vocabulary-free	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Average
	CLIP	✗	67.6	32.6	58.0	76.8	91.0	93.6	79.4	94.9	53.6	74.2	72.2
4-shot	SiM (images)	✓	67.0	39.5	79.6	70.7	84.5	80.7	95.5	95.0	59.1	80.3	75.2
	SiM (Wordnet)	✓	66.8	39.1	**82.6**	**73.2**	85.5	**87.8**	**96.3**	95.6	**62.5**	79.8	76.9
	SiM	✓	**67.2**	**43.0**	80.9	72.5	**85.6**	87.3	**96.3**	**95.8**	61.4	**81.8**	**77.2**
8-shot	SiM (images)	✓	71.7	47.3	84.5	76.9	87.5	87.4	97.0	97.3	66.5	83.0	79.9
	SiM (Wordnet)	✓	**71.8**	45.7	**87.1**	**79.8**	88.3	**91.7**	**98.3**	**96.9**	**70.0**	**83.8**	**81.4**
	SiM	✓	71.6	**47.6**	85.2	**79.8**	**88.4**	91.6	97.5	**96.9**	68.3	83.3	81.0
16-shot	SiM (images)	✓	74.0	**53.0**	**90.3**	82.9	89.1	91.7	98.5	96.8	71.3	**86.2**	83.4
	SiM (Wordnet)	✓	**74.7**	52.9	90.0	**84.1**	**89.7**	**93.8**	**98.9**	**97.5**	**74.2**	85.9	**84.2**
	SiM	✓	74.3	52.8	90.1	84.0	89.6	93.5	98.8	97.3	73.1	85.7	83.9

5. Results

Existing label mapping techniques are insufficient. Table 1 shows that state-of-the-art label mapping strategies consistently underperform compared to SiM across all 10 datasets. One-to-One mapping performs worse than BLM on average, highlighting the benefit of mapping multiple generic classes to each target category rather than relying on a rigid one-to-one correspondence. Surprisingly, the Centroids baseline—despite being a simple vision-only approach—largely outperforms BLM, raising questions about the viability of BLM in the few-shot learning setting for VLMs. When comparing BLM to SiM, we observe that Pets and Caltech101 exhibit the smallest performance gap, likely because their concepts are closely aligned with those in ImageNet. Conversely, Aircraft, Cars, and Flowers experience the largest performance drop, likely due to their fine-grained nature, which is either absent from ImageNet or only represented at a super-class level, making discrimination more challenging.

SiM is competitive with zero-shot classification. Table 1 presents zero-shot results, which are only achievable when class names are available. Our approach outperforms zero-shot CLIP on 6 out of 10 datasets with 4 shots per class and on 8 out of 10 datasets with 16 shots per class. However, performance varies significantly across datasets. On datasets such as Flowers, DTD, and EuroSAT, SiM surpasses zero-shot CLIP with only 4 shots per class, whereas for Food101 and Pets, SiM remains below zero-shot performance even in the 16-shot setting. These results suggest that incorporating even partial or noisy prior knowledge—such as an initial zero-shot prediction from a captioning model [9]—could further enhance SiM's performance, particularly in cases where vocabulary-free methods struggle to reach zero-shot accuracy.

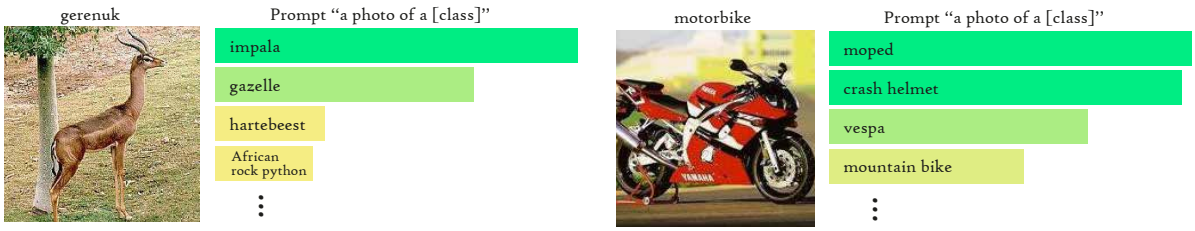
Table 3. Additional results on ImageNet. Top-1 accuracy averaged over 3 random seeds is reported. For SiM, embeddings $\boldsymbol{t}_k$ are obtained from a subset of 16452 words from Wordnet.

(a) Results for the CLIP ViT-B/16 backbone. For reference, the zero-shot (not vocabulary-free) performance is 66.7.

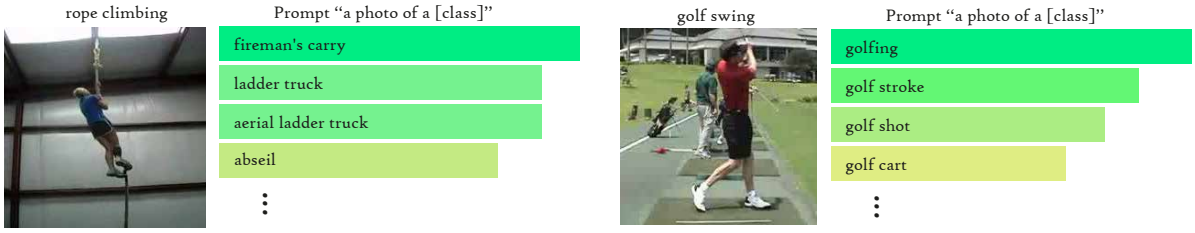
Method	4-shot	8-shot	16-shot	32-shot
Centroids	42.1	47.8	51.0	53.9
SiM (Wordnet)	**52.5**	**57.8**	**61.0**	**62.9**

(b) Results for the CLIP ViT-L/14 backbone. For reference, the zero-shot (not vocabulary-free) performance is 75.9.

Method	4-shot	8-shot	16-shot	32-shot
Centroids	50.9	57.2	60.7	63.2
SiM (Wordnet)	**63.9**	**68.7**	**71.2**	**72.8**



(a) With ImageNet text prompts ("a photo of a {class}") on two classes of the Caltech101 dataset (gerenuk and motorbike).



(b) With Wordnet text prompts ("a photo of a {class}") on two classes of the UCF101 dataset (rope climbing and golf swing).

(c) With ImageNet image prompts (one image per class) on two classes of the SUN397 dataset (nursery and shop bakery).

Figure 3. The linear mapping is learned in the 16-shot setting with the ViT-L/14 backbone. We order the generic prompts k according to the values of the weights $w_{k,c}$ for the given target class c, and retain the four highest. The heights of the bars are proportional to these values, while their color range from green (highest) to yellow (lowest).

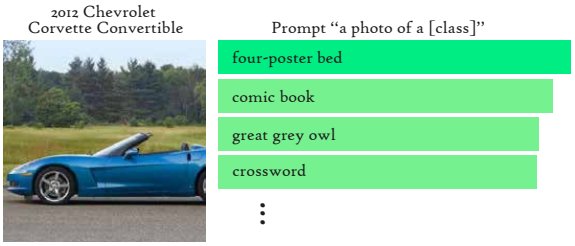
Lack of vocabulary degrades performances. Despite promising results, Table 1 shows that a performance gap remains between SiM and traditional few-shot learning methods, which benefit from explicit class names. However, this gap significantly decreases when using a more powerful backbone such as ViT-L/14, particularly for datasets like Flowers, Caltech101, and EuroSAT. We could also hypothesize that combining recent observations on current few-shot learning for VLMs with our approach (e.g., combination with prompt tuning, adapter or low-rank adaptation methods) could help bridge the remaining gap.

Images can be used as prompts but lag behind textual prompts. Table 2 show results obtained with different prompts. When the embeddings $\boldsymbol{t}_k$ are obtained from a single image from the classes of ImageNet, performances tend to be slightly lower than when using ImageNet class names. This gap narrows when the number of shots increases, or when using a more powerful backbone. In the 16-shot setting with ViT-L/14, image-based and text-based prompts yield comparable performance on most datasets.

Textual prompts are not restricted to ImageNet classes. Table 2 show results obtained with different prompts, namely a subset of $K = 16452$ words from Wordnet to generate the embeddings. Our approach scales gracefully as learning the mapping with 16-shot ImageNet ($N = 16000$) takes 0.8s on a Tesla A100. The performances slightly degrade compared to the ImageNet classes in the 4-shot setting for both backbones. This could be due to an increased sensitiveness to noise because of the much larger K, where the learned mapping generalizes not as effectively at test time. In the 8 and 16-shot settings, performances are on par on average. In this setting, we can also provide results for ImageNet, presented in Table 3. Similar to some other datasets, e.g. Food, the performance of SiM remains below zero-shot even with 16 and 32 shots. The gap is narrower with the more powerful ViT-L/14 backbone.

Learned weights can provide semantic understanding. Beyond classification performance, we can investigate the relative importance of weights $w_{k,c}$ for a given target class c (see Eq. (5)) which potentially provide a high level seman-

(a) An ambiguous example from the EuroSAT dataset with target class ”Highway or road”, linked to semantically related concepts which cannot be seen on the visual shots.



(b) An ambiguous example from the StanfordCars dataset with target class ”2012 Chevrolet Corvette Convertible”, for which no meaningful generic prompt can be matched.

Figure 4. The linear mapping is learned in the 16-shot setting with the ViT-L/14 backbone. We order the generic prompts k according to the values of the weights $w_{k,c}$ for the given target class c, and retain the four highest. The heights of the bars are proportional to these values, while their color range from green (highest) to yellow (lowest).

tic understanding. Figure 3 illustrates this interpretability across the three types of prompts studied in this work. Figure 3a presents results for ImageNet class names prompts on Caltech101 (animals and objects), where our approach meaningfully associates gerenuk with impala and gazelle, other antelope species. Figure 3b shows results for the UCF101 dataset (action recognition) using Wordnet-based prompts, where semantically related words to rope climbing such as abseil appear as relevant matches. Figure 3c presents results for the SUN397 dataset (scene recognition) using image-based prompts, demonstrating that a shop bakery is linked to baked goods such as bagels, dough, and pretzels.

The linear mapping is not always interpretable. The information contained in the learned mapping is not straightforwardly interpretable for every target classes. For instance, Figure 4a shows an example where the most important weights are associated with generic prompts semantically related to the target class, but which do not appear in the images defining it. An other interpretability failure case appears to be linked with fine-grained datasets, as shown in Figure 4b, where the target class is matched with completely unrelated concepts. Interestingly, this lack of interpretability does not seem to be detrimental to accuracy, as SiM achieves a 31% reduction of the error rate in the 16-shot setting with the ViT-L/14 backbone (i.e., the setting with which Figure 4b was obtained) compared to zero-shot classification with class names. Therefore, the choice of generic prompts may induce a trade-off between interpretability and separability of target classes.

6. Discussion and future works

As only few works have been targeting the problem of vocabulary-free image classification so far, the avenues for future research seem numerous in vocabulary-free few-shot learning for VLMs. The results presented in Table 1 suggest there may be opportunities for adapting recent advances in few-shot adaptation for VLM, such as prompt tuning, to this novel setting. Furthermore, our baseline could be tweaked by proposing a different regularization, to promote higher classification accuracy or better interpretability.

Beyond refining the learning algorithm, enhancing the choice of generic prompts could also play a crucial role in improving performance. As seen in Table 2, the set of generic prompts may come from very diverse sources, suggesting potential for optimizing the choice of the embeddings $\boldsymbol{t}_k$ used to compute the similarities. This in turn could be helpful to mitigate issues illustrated in Figure 4b.

Finally, associating meaningful names with groups of shots remains an open challenge. For example, by investigating techniques for automatically labeling discovered classes while evaluating performance with semantic metrics, such as semantic intersection over union (IoU), as suggested in [9].

7. Conclusion

In this work, we introduced *vocabulary-free few-shot learning*, a new framework for image classification using VLMs where class names are not available. We highlighted how current few-shot adaptation methods are ill-equipped to handle this practical scenario. To address this limitation, we proposed a simple yet effective baseline that leverages similarity scores between few-shot images and a set of generic, arbitrary prompts, which can be sourced from texts or images, predefined without any knowledge of the target classes. Interestingly, our least-squares baseline, SiM, achieves performance close to that of more complex few-shot adaptation techniques that rely on explicit class names. Additionally, SiM does not require direct access to the embeddings of either the generic prompts or the images—only the similarity scores. Finally, we demonstrated that analyzing the relative weights of the learned linear mapping could potentially provide high-level semantic insights into the target classes. Overall, we hope this work serves as a stepping stone for *vocabulary-free few-shot learning*, an important yet overlooked problem in VLM adaptation.

8. Acknowledgments

M. Zanella is funded by the Walloon region under grant No. 2010235 (ARIAC by DIGITALWALLONIA4.AI). C. Fuchs is funded by the MedReSyst project, supported by FEDER and the Walloon Region. Part of the computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under Grant No. 2.5020.11 and by the Walloon Region.

References

[1] Christian Bauckhage. K-means clustering is matrix factorization. *arXiv preprint arXiv:1512.07548*, 2015. 4

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 446–461. Springer, 2014. 4

[3] Adrian Bulat and Georgios Tzimiropoulos. Lasp: Text-to-text optimization for language-aware soft prompting of vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23232–23241, 2023. 2

[4] Chengyi Cai, Zesheng Ye, Lei Feng, Jianzhong Qi, and Feng Liu. Bayesian-guided label mapping for visual reprogramming. *Advances in Neural Information Processing Systems*, 37:17656–17695, 2025. 3, 4

[5] Aochuan Chen, Yuguang Yao, Pin-Yu Chen, Yihua Zhang, and Sijia Liu. Understanding and improving visual prompting: A label-mapping perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19133–19143, 2023. 2, 3

[6] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Plot: Prompt learning with optimal transport for vision-language models. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 4

[8] Alessandro Conti, Enrico Fini, Massimiliano Mancini, Paolo Rota, Yiming Wang, and Elisa Ricci. Vocabulary-free image classification. In *Advances in Neural Information Processing Systems*, pages 30662–30680. Curran Associates, Inc., 2023. 2

[9] Alessandro Conti, Enrico Fini, Massimiliano Mancini, Paolo Rota, Yiming Wang, and Elisa Ricci. Vocabulary-free image classification and semantic segmentation. *arXiv preprint arXiv:2404.10864*, 2024. 2, 6, 8

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 4

[11] Mohammad Mahdi Derakhshani, Enrique Sanchez, Adrian Bulat, Victor Guilherme Turrisi da Costa, Cees GM Snoek, Georgios Tzimiropoulos, and Brais Martinez. Variational prompt tuning improves generalization of vision-language models. *arXiv preprint arXiv:2210.02390*, 2022. 2

[12] Gamaleldin F Elsayed, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial reprogramming of neural networks. *arXiv preprint arXiv:1806.11146*, 2018. 2, 3

[13] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 4

[14] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, pages 1–15, 2023. 2

[15] Gene H Golub, Per Christian Hansen, and Dianne P O'Leary. Tikhonov regularization and total least squares. *SIAM journal on matrix analysis and applications*, 21(1):185–194, 1999. 3

[16] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 4

[17] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 1, 2, 4

[18] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15190–15200, 2023. 2

[19] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 4

[20] Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton van den Hengel. Retrieval augmented classification for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6959–6969, 2022. 2

[21] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5206–5215, 2022. 2

[22] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 4

[23] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 2, 4

[24] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 4

[25] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 4

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1

[27] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 2

[28] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 2

[29] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15638–15650, 2022. 2

[30] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 4

[31] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 4

[32] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 2

[33] Yun-Yun Tsai, Pin-Yu Chen, and Tsung-Yi Ho. Transfer learning without knowing: Reprogramming black-box machine learning models with scarce data and limited resources. In *International Conference on Machine Learning*, pages 9614–9624. PMLR, 2020. 2, 3, 4

[34] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 4

[35] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6757–6767, 2023. 2

[36] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10899–10909, 2023. 1, 2, 4

[37] Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1593–1603, 2024. 1, 2, 4

[38] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision*, pages 493–510. Springer, 2022. 1, 2, 4

[39] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. 2

[40] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 2, 4

[41] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15659–15669, 2023. 1, 2