Test-time Adaptation of Medical Vision-Language Models

Fereshteh Shakeri[1,2⋆], Ghassen Baklouti[1,2*], Julio Silva-Rodríguez[1], Maxime Zanella[3], Houda Bahig[2], Jose Dolz[1,2], and Ismail Ben Ayed[1,2]

[1]ETS Montreal, Canada [2]CRCHUM, Canada [3]UCLouvain, Belgium

Abstract. Integrating image and text data through multi-modal learning has gained significant attention in medical imaging research, building on its success in computer vision. While substantial progress has been made in developing medical foundation models and enabling their zero-shot transfer to downstream tasks, the potential of test-time adaptation remains largely underexplored in this domain. Inspired by recent advances in transductive learning and parameter-efficient fine-tuning, we investigate test-time adaptation of medical vision-language models. Our method leverages information-theoretic principles, maximizing the mutual information between the visual inputs and the text-based class representations, while minimizing a Kullback-Leibler divergence term penalizing deviation of the predictions from the zero-shot outputs. Building on this foundation, we introduce the first structured benchmark for test-time adaptation of medical vision-language models, exploring strategies tailored to the unique challenges of medical imaging. Our extensive experiments include two medical modalities, three specialized foundation models, six downstream tasks, and multiple state-of-the-art test-time adaptation methods, demonstrating significant performance improvements and establishing a new benchmark for this emerging field. The code is available at: https://github.com/FereshteShakeri/TTAMedVLMs.

Keywords: Medical Vision-Language Models · Test Time Adaptation · Transductive Inference · Parameter-Efficient Fine-tuning

1 Introduction

Deep neural networks have gained significant attention over the past decade in the medical image analysis community [19], driven by their transformative success in natural image recognition. Their advancements have been effectively applied to various medical imaging tasks, including tumor grading in gigapixel-stained histology images [30], diabetic retinopathy classification [7], and radiology image interpretation [13], among others. However, despite their potential, the adoption of these models in real-world clinical settings remains limited. A major challenge lies in their dependence on large, annotated datasets, which are often difficult to obtain in medical domains [5]. Additionally, the presence

⋆ Equal contributions. Corresponding author: fereshteh.shakeri.1@etsmtl.net

of significant domain shifts—arising from inter-scanner variability, staining differences, and population heterogeneity—necessitates continuous adaptation for reliable performance across diverse clinical environments. A promising solution to address these challenges is transfer learning, where pre-trained models are leveraged to extract robust feature representations. While widely successful in computer vision, direct transfer from natural to medical images has not yielded the anticipated improvements [27], largely due to the fine-grained nature of medical imaging data and the unique complexities inherent to clinical datasets.

The field of transfer learning is undergoing a major transformation, driven by the emergence of foundation models that leverage large-scale pre-training on diverse datasets to enhance generalization across multiple tasks. Among these, vision-language models (VLMs) such as CLIP [26] and ALIGN [14] have demonstrated exceptional adaptability, utilizing contrastive learning to align visual and textual representations at scale. These models exhibit strong robustness to domain shifts [26], making them particularly effective for zero-shot learning and adaptable to low-shot settings or inference-time adjustments. However, despite their success in natural image domains, their direct application to medical imaging has been severely limited, primarily due to the lack of fine-grained medical expertise encoded within these models. Unlike general-purpose datasets, medical data requires specialized clinical knowledge, which current VLMs fail to capture, leading to suboptimal performance in real-world clinical settings.

To address this gap, significant efforts have been made to develop domain-specific medical VLMs, leveraging large open-access medical datasets across various specialties, including histology [11, 12, 21], ophthalmology [8, 32, 39], and radiology [35, 36, 45]. While these models provide a strong foundation, their adaptation to downstream medical tasks remains a challenge. Some studies have explored few-shot adaptation, either by training on a reduced percentage of the available datasets (e.g., 1% or 10% of samples in [12, 45])—which still requires hundreds or thousands of annotated examples—or by establishing few-shot learning benchmarks [28] aligned with standard adaptation protocols. However, few-shot learning still requires labeled data, which may not always be available in real-world clinical scenarios. In contrast, test-time adaptation (TTA) offers a more practical and scalable alternative, allowing models to dynamically adapt to new domains without the need for labeled target data [34]. By adjusting model parameters [25] or refining predictions [42] based on incoming test samples, TTA can mitigate distributional shifts at inference time, making it particularly suitable for medical imaging applications, where domain variability is a persistent challenge. While TTA has gained increasing attention in computer vision, with recent and early explorations of adapting VLMs for natural image classification [25, 42, 44], its potential for medical imaging remains largely unexplored.

In this work, we introduce the first structured benchmark for test-time adaptation of medical VLMs, systematically evaluating adaptation techniques across multiple modalities and datasets. Inspired by recent advances in parameter-efficient fine-tuning (PEFT) and transductive learning, we investigate a novel language-aware test-time adaptation method. The approach maximizes the mu-

tual information between the visual inputs and the text-based class representations, while preserving the model's zero-shot capabilities through a Kullback-Leibler divergence regularization. Our contributions could be summarized as follows:

- We present the first structured test-time adaptation benchmark for medical vision-language models, including two medical imaging domains: histology, and ophthalmology.
- We evaluate a range of test time adaptation strategies, including entropy minimization and transductive learning approaches, providing a comprehensive comparison of state-of-the-art TTA methods.
- We investigate a novel language-aware adaptation framework that enhances test-time adaptation by maximizing the mutual information between the input images and the text-based class descriptions, while preserving the zero-shot knowledge.
- We conduct extensive experiments on six downstream tasks using specialized foundation models, demonstrating that our language-aware TTA method consistently outperforms existing adaptation techniques.

2 Related Work

Medical vision-language models. Building large-scale medical vision-language datasets is significantly more challenging than in natural image domains, where models like CLIP [26] and ALIGN [14] are trained on hundreds of millions of image-text pairs. In medical imaging, privacy regulations, specialized expertise, and limited annotations often constrain dataset availability. To address this, recent efforts have focused on developing vision-language models specifically for medical applications, spanning various domains such as radiology [35, 36, 45], ophthalmology [32, 39], and computational pathology [11, 12]. Notably, Quilt-1M [12], one of the largest vision-language histopathology datasets, aggregates one million image-text pairs from online sources, while ViLReF [39] is pretrained on over 450,000 retinal images with the corresponding diagnostic reports.

Test-time adaptation of vision-only models. Adapting models to new incoming data at test time has gained significant attention, with one of the initial works, TENT [34], which minimizes the entropy of predictions by updating the batch-norm statistics. Afterwards, LAME [4] highlighted the problem of model collapse when facing class imbalance, and proposes to only adapt the predictions, using Laplacian regularization. SAR [24] improves TENT with a sharpness-aware entropy loss, encouraging convergence to flatter loss regions and improving generalization. While these techniques primarily focused on natural-image tasks, they have also inspired research in medical imaging [2, 10, 16, 20, 33, 38]. While these advances in test-time adaptation can still bring significant improvements to medical vision-language models, tailored methods benefiting from the text modality are needed to fully leverage the adaptation capacity of these models.

Test-time adaptation for VLMs. Following the success of test-time adaptation in vision-only models and the remarkable zero-shot capabilities of CLIP,

some approaches have emerged to adapt vision-language models at test time, but their application remains limited to natural images. One key distinction among these methods lies in the parameters they modify, ranging from input prompts [29] and intermediate layers, as in WATT [25], to adapters serving as memory banks [17] and training-free strategies [43]. A second distinction lies in how they process the incoming data. In this work, we focus on the setting where predictions must be made independently for each batch. This setup is particularly well-suited for transductive learning. For instance, TransCLIP [41, 42] models each class with a multivariate Gaussian distribution, while enforcing consistency between the predicted labels and zero-shot predictions.

3 Methods

Zero-shot classification. At inference time, a vision language model performs zero shot classification by leveraging a simple yet effective text-image alignment strategy to generate final predictions. Given a set of k candidate classes, textual descriptions, a.k.a prompts, are created for each class. For instance, if one of the classes is *cancer*, the corresponding prompt might be [c_k = *a histopathology slide showing cancer cell.*]. These prompts are then processed through the text encoder θ_t, producing the corresponding text embeddings $t_k = \theta_t(c_k)$. Simultaneously, each test image x_i is passed through the vision encoder θ_v, yielding the visual embedding $f_i = \theta_v(x_i)$. The posterior probability of class k given input test image x_i is given by:

$$p_{i,k} = \frac{\exp(l_{i,k}/\tau)}{\sum_j^K \exp(l_{i,j}/\tau)} \tag{1}$$

where τ is a temperature scaling parameter and $l_{i,k} = f_i^\top t_k$ is the logit score derived from the cosine similarity between the visual and text embeddings.

Proposed objective for TTA. To further enhance the zero-shot predictions at inference time, we minimize a language-aware information maximization objective function tailored for the test time adaptation of VLMs, which we introduced recently in the context of natural-image few-shot tasks [1]:

$$-I(\theta_v, \theta_t) + T(\theta_v, \theta_t) \tag{2}$$

In the following, we describe each of the two terms of the objective function above. $I(\theta_v, \theta_t)$ represents the mutual information between the vision inputs and the text-based class descriptions, considering them both as random variables, denoted as $\mathcal{X}_v$ and $\mathcal{X}_t$, respectively. It integrates conditional entropy $H(\mathcal{X}_t|\mathcal{X}_v; \theta_v, \theta_t)$, which minimizes the uncertainty associated with the text-based prediction of each test sample, and marginal entropy $H(\mathcal{X}_t; \theta_v, \theta_t)$, which promotes a balanced class distribution, ensuring diversity in class predictions and preventing the model from favoring a limited number of classes. Formally, this

reads:

$$\begin{aligned} I(\theta_v, \theta_t) &= \lambda_{\text{ent}} H(\mathcal{X}_t) - \lambda_{\text{cond}} H(\mathcal{X}_t|\mathcal{X}_v) \\ &= -\lambda_{\text{ent}} \sum_{k=1}^{K} \bar{p}_k \log \bar{p}_k + \lambda_{\text{cond}} \sum_{i \in B} \sum_{k=1}^{K} p_{i,k} \ln p_{i,k} \end{aligned} \tag{3}$$

where λ_{ent} and λ_{cond} are two non-negative weighting factors that control the contribution of each term, and $\bar{p}_k = \frac{1}{|B|} \sum_{i \in B} p_{i,k}$ is the marginal probability of class k over the batch of test samples. B is the set of indices of the test samples in the batch. These entropy terms work together to promote confident and class-balanced predictions of the unlabeled test samples. In addition, regularization term $T(\theta_v, \theta_t)$ provides text-based supervision to preserve the model's zero-shot capabilities. It penalizes deviation of the model's predictions, $\mathbf{p}_i = (p_{i,k})_{1 \leq k \leq k}$, from the zero-shot ones, $\hat{\mathbf{y}}_i = (y_{i,k})_{1 \leq k \leq k}$. This is done by minimizing the following weighted Kullback-Leibler (KL) divergence:

$$T(\theta_t, \theta_v) = \lambda_{\text{txt}} \sum_{i \in B} \text{KL}(\mathbf{p}_i || \hat{\mathbf{y}}_i) = \sum_{i \in B} \mathbf{p}_i^\top \log \frac{\mathbf{p}_i}{\hat{\mathbf{y}}_i} \tag{4}$$

Also, similarly to the entropy terms, the text regularization term is weighted by a factor λ_{txt} to control its influence within our overall loss.

Fine-tuning. To efficiently optimize the proposed loss, we adopt LoRA [9,40] as a fine-tuning strategy. This amounts to approximating the incremental updates ΔW of the original pre-trained weight matrix $W_0 \in \mathrm{R}^{m \times n}$ as the product of two low-rank matrices, A and B, based on the concept of ''intrinsic rank" of a downstream task. Specifically, the low-rank weight modification is expressed as:

$$W = W_0 + \Delta W = W_0 + BA \tag{5}$$

where $A \in \mathrm{R}^{r \times n}$ and $B \in \mathrm{R}^{m \times r}$, with r representing the intrinsic rank, which is typically much smaller than m and n (i.e., $r << (m, n)$).

4 Experiments

4.1 Setup

Medical VLMs. Two predominant medical imaging modalities are included: histology, and ophthalmology. In each application, specialized VLMs that underwent modality-specific pre-training through textual supervision are leveraged. **Histology**: we employed Quilt-1M [12], which utilizes ViT-B/32 and ViT-B/16 vision encoder and a GPT-2 text encoder. **Ophthalmology**: we adopt ViLReF [39], a foundation model based on Chinese CLIP (CN-CLIP) [37], which has been pre-trained on several large-scale datasets containing approximately 200 million image-text pairs. ViLReF employs ViT-B/16 as its vision encoder and RoBERTa-wwm-ext-base-Chinese [6] as its text encoder.

Adaptation tasks. We include a diverse set of tasks for TTA, designed to evaluate the transfer capabilities of medical VLMs. Each foundation model is applied to datasets within its respective medical domain. Note that all datasets are carefully selected to prevent test data leakage, ensuring that no samples used for evaluation were involved in the model's pre-training. In histology, the evaluation covers three different organs and cancer types, including colorectal adenocarcinoma (NCT-CRC) [18] , prostate cancer grading (SICAP-MIL) [31], and Lung cancer (LC25000-Lung) [3]. For ophthalmology, MESSIDOR [7] is focused on diabetic retinopathy (DR) grading. Also, FIVES [15] and ODIR200x3 [23], are incorporated to the benchmark, addressing inter-disease discrimination.

TTA protocol and evaluation. We conduct test-time adaptation following the standard practices established in vision literature [25]. A fixed batch size of 128 is employed throughout all experiments to ensure consistency and facilitate fair comparisons across different scenarios. Each batch is processed independently, and after adaptation and evaluation, the model is reset to prevent any information leakage between batches. Model performance is assessed using average class-wise accuracy (ACA), a widely used metric in medical imaging benchmarks [22]. All results are averaged over three random seeds across all experiments to ensure robustness and account for variability.

Implementation details. The proposed method is deployed using a fixed LoRA rank of 2. For each input batch, 10 adaptation steps are carried out, updating the visual encoder's selected parameters. ADAM is the optimizer, using a learning rate of 10^{-3}. The marginal entropy weight is set to $\lambda_{\text{ent}} = 5$, the conditional entropy to $\lambda_{\text{cond}} = 1$ and the text regularization term is $\lambda_{\text{txt}} = 0.1$.

Baselines. A comprehensive evaluation of current TTA methods is carried out. Three different types of baselines are included. *i*) Zero-shot [26], i.e., no adaptation, where predictions are derived by computing the softmax cosine similarity between image and text embeddings. Text embeddings are generated using the specific prompts established in each medical VLMs. In particular, we employ prompt ensembles for Quilt-1M [12] and ViLReF [39], leveraging domain-specific textual representations to enhance alignment with visual features. *ii*) Black-box TTA methods that rely uniquely on embedding representation. Transductive approaches LAME [4], and TransCLIP [42] are incorporated, whose hyper-parameters are set as in the original publications. *iii*) Methods updating intermediate layers of vision or language encoders. We incorporate an adapted version of TENT [34] in which only the affine parameters in Layer Normalization (LN) layers are updated, applying entropy minimization directly to the model's logits. Notably, we perform 10 iterations of TENT adaptation per batch. We employ Adam optimizer with a fixed learning rate of 10^{-3}. Keeping this same learning configuration, we include SAR [24] and WATT [25]. We focus on WATT-p version, which optimizes the TTA loss separately for different models ($H = 4$ for histology, $H = 10$ for retina), each utilizing distinct textual templates.

Table 1: **Comparison of state-of-the-art TTA methods.**

	Zero-Shot [26] ICLR'21	TENT [34] ICLR'21	SAR [24] ICCV'23	LAME [4] CVPR'22	TransCLIP [42] NeurIPS'24	WATT [25] NeurIPS'24	Ours
(a) Histology - Quilt-B/32 [12]							
NCT-CRC	52.62	60.33	60.12	60.57	58.56	59.96	**61.99**
LC-LUNG	80.57	79.37	80.73	85.68	**92.73**	87.52	89.02
SICAP-MIL	27.51	27.25	27.56	23.12	17.28	29.38	**29.45**
Avg.	53.57	55.65	56.14	56.46	56.19	58.95	**60.15**
(b) Histology - Quilt-B/16 [12]							
NCT-CRC	29.90	27.65	30.40	31.72	36.90	**49.67**	43.93
LC-LUNG	56.70	51.90	57.20	59.05	54.25	81.82	**85.98**
SICAP-MIL	32.78	28.40	33.60	28.54	48.26	43.90	**49.30**
Avg.	39.79	35.98	40.40	39.77	46.47	58.46	**59.74**
(c) Ophtalmology - ViLReF [39]							
MESSIDOR	42.66	39.31	36.29	45.82	28.89	26.88	**48.92**
FIVES	75.00	69.84	70.37	74.32	75.62	**75.99**	74.25
ODIR200x3	68.47	70.90	72.19	68.06	**87.43**	86.87	84.17
Avg.	62.04	60.02	59.62	62.73	63.98	63.25	**69.11**

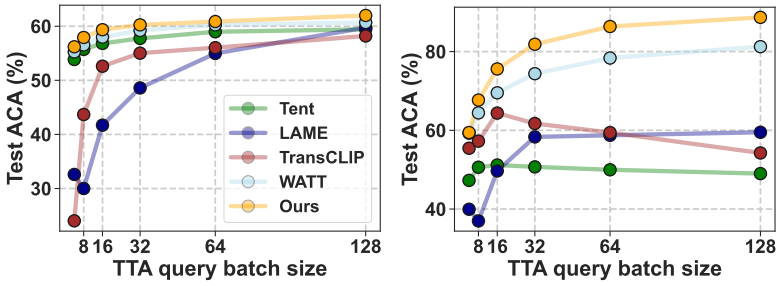
4.2 Results

Performance analysis. Table 1 provides a comprehensive comparison of our method against state-of-the-art TTA baselines in the histology and ophthalmology domains. ***i*) Comparison to black-box approaches**, i.e., LAME and TransCLIP. Our method achieves superior average performance across both modalities, significantly outperforming such baselines. TransCLIP, which operates without fine-tuning model parameters, performs well in cases where the number of classes is small and the zero-shot performance is already strong (e.g., three-class datasets such as LC-Lung and ODIR200x3). However, as the number of classes increases or when zero-shot performance is weaker, its effectiveness declines since it lacks the expressiveness of our proposed approach. This can be observed in SICAP-MIL with the Quilt-B32 model, where TransCLIP's performance drops by 10% compared to the zero-shot baseline. ***ii*) Comparison to parameter-efficient approaches**: TENT, SAR, and WATT. Our method outperforms TENT and SAR by large margins. For example, for histology tasks using Quilt-B/16 the average improvements are substantial: +23.8% and +19.3% respectively. Compared to WATT, our approach provides similar performance for histology tasks, but average improvements of +5.9% in ophthalmology. It is worth noting that while prior works show sometimes degrade its performance during TTA in at least one dataset, *the proposed method is the most robust, not suffering large degradations w.r.t. zero-shot predictions.*

Computational efficiency. The results in Table 2 highlight the computational trade-offs between different adaptation strategies. LAME and TransCLIP exhibit the fastest evaluation times ($\sim$ 1.3 seconds), as they rely on lightweight optimization strategies without modifying model parameters. In contrast, TENT and SAR require slightly longer processing times since they require backprop-

Table 2: **Computational workload.** Runtime comparison of TTA methods for processing a batch of 128 samples (LC-LUNG dataset). All experiments were conducted on a single NVIDIA GeForce RTX 3090 (24GB) GPU.

	Black-box adaptation		Parameter-Efficient Fine-Tuning			
	LAME [4]	TransCLIP [42]	TENT [34]	SAR [24]	WATT [25]	*Ours*
Runtime (sec.)	∼ 1.3	∼ 1.3	∼ 12.4	∼ 13.5	∼ 750.0	∼ 15.1



(a) Quilt-B/32 - NCT **(b) Quilt-B/16 - LUNG**

Fig. 1: **Ablation study.** Effect of query batch size in TTA.

agation through internal parameters in the vision encoder during adaptation. While achieving stronger adaptation performance, WATT incurs a significantly higher computational cost (∼ 12.4 minutes), making it impractical for real-time applications. Our proposed method is computationally closer to TENT or SAR, maintaining a reasonable runtime, yet yielding outstanding results, even surpassing WATT. Overall, *our method demonstrates state-of-the-art performance while maintaining a balance between effectiveness and efficiency*, making it a compelling choice for test-time adaptation in medical vision-language models.

Robustness to query batch size. Fig. 1 illustrates the robustness of the different TTA methods when receiving data in smaller batches. Transductive strategies such as TransCLIP may suffer severe performance drops when accessing only small data batches since they cannot leverage the whole data distribution. In contrast, parameter-efficient strategies such as WATT and our proposed method are the more robust solution, providing satisfactory performance even when receiving query batches of 32 images. Considering the different modalities, our approach is the most stable solution for this challenging study.

5 Conclusions

We introduced the first structured benchmark for test-time adaptation (TTA) of medical vision-language models (VLMs), systematically evaluating a range of adaptation strategies across multiple domains. In addition, we investigated a

novel language-aware test-time adaptation framework that maximizes the mutual information between visual inputs and textual class representations, while preserving the model's zero-shot capabilities. Extensive experiments on histology and ophthalmology data showed that our method consistently outperforms existing adaptation techniques. Additionally, our computational analysis highlights that this does not come at the price of a high computational cost. Our findings underscore the importance of using both visual and textual modalities for more robust adaptation, paving the way for future research on test-time adaptation for medical VLMs.

Acknowledgments. This work was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) and Montreal University Hospital Research Center (CRCHUM). We also thank Calcul Quebec and Compute Canada.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article

References

1. Baklouti, G., Zanella, M., Ayed, I.B.: Language-aware information maximization for transductive few-shot clip (2025) 4
2. Bateson, M., Lombaert, H., Ben Ayed, I.: Test-time adaptation with shape moments for image segmentation. In: MICCAI. pp. 736–745 (2022) 3
3. Borkowski, A.A., et al.: Lung and colon cancer histopathological image dataset (lc25000). arXiv preprint arXiv:1912.12142 (2019) 6
4. Boudiaf, M., Mueller, R., Ben Ayed, I., Bertinetto, L.: Parameter-free online test-time adaptation. In: CVPR. pp. 8344–8353 (2022) 3, 6, 7, 8
5. Chen, X., et al.: Recent advances and clinical applications of deep learning in medical image analysis. Medical Image Analysis **79** (2022) 1
6. Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., Hu, G.: Revisiting pre-trained models for chinese natural language processing. In: EMNLP. pp. 657–668 (2020) 5
7. Decencière, E., et al.: Feedback on a publicly distributed image database: The messidor database. Image Analysis & Stereology **33**, 231–234 (07 2014) 1, 6
8. Du, J., et al.: Ret-clip: A retinal image foundation model pre-trained with clinical diagnostic reports. In: MICCAI. pp. 709–719 (2024) 2
9. Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: LoRA: Low-rank adaptation of large language models. In: ICLR (2021) 5
10. Hu, M., Song, T., Gu, Y., Luo, X., Chen, J., Chen, Y., Zhang, Y., Zhang, S.: Fully test-time adaptation for image segmentation. In: MICCAI. pp. 251–260 (2021) 3
11. Huang, Z., et al.: A visual–language foundation model for pathology image analysis using medical twitter. Nature Medicine **29**, 1–10 (2023) 2, 3
12. Ikezogwo, W., Seyfioglu, S., Ghezloo, F., Geva, D., Sheikh Mohammed, F., Anand, P.K., Krishna, R., Shapiro, L.: Quilt-1M: One million image-text pairs for histopathology. NeurIPS **36**, 37995–38017 (2023) 2, 3, 5, 6, 7
13. Irvin, J., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: AAAI. vol. 33, pp. 590–597 (2019) 1
14. Jia, C., et al.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML. pp. 4904–4916 (2021) 2, 3

15. Jin, K., Huang, X., Zhou, J., Li, Y., Yan, Y., Sun, Y., Zhang, Q., Wang, Y., Ye, J.: Fives: A fundus image dataset for artificial intelligence based vessel segmentation. Scientific Data **9**, 475 (08 2022) 6
16. Karani, N., et al.: Test-time adaptable neural networks for robust medical image segmentation. Medical Image Analysis **68**, 101907 (2021) 3
17. Karmanov, A., Guan, D., Lu, S., El Saddik, A., Xing, E.: Efficient test-time adaptation of vision-language models. In: CVPR. pp. 14162–14171 (2024) 4
18. Kather, J.N., Halama, N., Marx, A.: 100,000 histological images of human colorectal cancer and healthy tissue. Zenodo10 **5281** (2018) 6
19. Litjens, G., et al.: A survey on deep learning in medical image analysis. Medical Image Analysis (2017) 1
20. Liu, Q., et al.: Single-domain generalization in medical image segmentation via test-time adaptation from shape dictionary. In: AAAI. vol. 36, pp. 1756–1764 (2022) 3
21. Lu, M.Y., et al.: Visual language pretrained multiple instance zero-shot transfer for histopathology images. In: CVPR. pp. 19764–19775 (2023) 2
22. Maier-Hein, L., et al.: Metrics reloaded: recommendations for image analysis validation. Nature Methods **21** (2024) 6
23. NIHDS-PKU: Competition on Ocular Disease Intelligent Recognition (ODIR). https://odir2019.grand-challenge.org (2019) 6
24. Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., Tan, M.: Towards stable test-time adaptation in dynamic wild world. In: ICLR (2023) 3, 6, 7, 8
25. Osowiechi, D., et al.: Watt: Weight average test time adaptation of clip. NeurIPS **37**, 48015–48044 (2025) 2, 4, 6, 7, 8
26. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021) 2, 3, 6, 7
27. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: Understanding transfer learning for medical imaging. In: NeurIPS. pp. 1–11 (2019) 2
28. Shakeri, F., et al.: Few-shot adaptation of medical vision-language models. In: MICCAI. pp. 553–563 (2024) 2
29. Shu, M., Nie, W., Huang, D.A., Yu, Z., Goldstein, T., Anandkumar, A., Xiao, C.: Test-time prompt tuning for zero-shot generalization in vision-language models. NeurIPS **35**, 14274–14289 (2022) 4
30. Silva-Rodríguez, J., et al.: Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. Computer methods and programs in biomedicine **195**, 105637 (2020) 1
31. Silva-Rodríguez, J., et al.: Proportion constrained weakly supervised histopathology image classification. Computers in Biology and Medicine **147**, 105714 (2022) 6
32. Silva-Rodriguez, J., et al.: A foundation language-image model of the retina (FLAIR): Encoding expert knowledge in text supervision. Medical Image Analysis **99**, 103357 (2025) 2, 3
33. Valanarasu, J.M.J., et al.: On-the-fly test-time adaptation for medical image segmentation. In: MIDL. pp. 586–598. PMLR (2024) 3
34. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: ICLR (2021) 2, 3, 6, 7, 8
35. Wang, Z., et al.: Medclip: Contrastive learning from unpaired medical images and text. In: EMNLP. pp. 1–12 (10 2022) 2, 3
36. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In: ICCV (2023) 2, 3

37. Yang, A., et al.: Chinese clip: Contrastive vision-language pretraining in chinese. arXiv preprint arXiv:2211.01335 (2022) 5
38. Yang, H., et al.: DLTTA: Dynamic learning rate for test-time adaptation on cross-domain medical images. IEEE TMI **41**(12), 3575–3586 (2022) 3
39. Yang, S., et al.: Vilref: An expert knowledge enabled vision-language retinal foundation model. IEEE TMI (2024) 2, 3, 5, 6, 7
40. Zanella, M., Ben Ayed, I.: Low-rank few-shot adaptation of vision-language models. In: CVPRw. pp. 1593–1603 (2024) 5
41. Zanella, M., et al.: Boosting vision-language models for histopathology classification: Predict all at once. In: MedAGI MICCAIw. pp. 153–162 (2024) 4
42. Zanella, M., et al.: Boosting vision-language models with transduction. NeurIPS **37**, 62223–62256 (2024) 2, 4, 6, 7, 8
43. Zanella, M., et al.: On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning? In: CVPR. pp. 23783–23793 (2024) 4
44. Zanella, M., et al.: Realistic test-time adaptation of vision-language models. In: CVPR. pp. 25103–25112 (2025) 2
45. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: MHLC (2022) 2, 3