CANDIDATE-CONDITIONED ALIGNMENT FOR STABLE TEST-TIME ADAPTATION IN CROSS-DOMAIN VQA

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) have achieved remarkable progress in visual question answering (VQA), yet their performance often degrades under domain shifts. Test-time domain adaptation (TTDA) offers an attractive solution by adapting models directly on unlabeled target data during inference. Existing TTDA methods, however, typically optimize prediction confidence over a shared answer space, an assumption that is inconsistent with the question-dependent prediction structure of VQA. As a result, optimization may progressively become biased toward dominant answers and produce question-insensitive predictions under domain shift. To address this issue, we propose a Candidate-Conditioned Alignment framework for VQA test-time domain adaptation. Instead of adapting over the global answer space, our method performs adaptation within instance-specific candidate sets through a local energy-based image-text alignment objective, enabling stable optimization while preserving question-conditioned reasoning. A lightweight text adapter further supports efficient adaptation without updating the pretrained backbone. Extensive experiments on six vision-language models and six cross-domain VQA benchmarks demonstrate consistent improvements across diverse domains, achieving an average accuracy gain of **18.8%** over direct candidate-based inference under the same candidate setting.

1 INTRODUCTION

Visual Question Answering (VQA) requires joint reasoning over visual content and natural language questions. Recent vision-language models (VLMs) Zhao et al. (2023); Zhang et al. (2024a); Zhou et al. (2022), such as CLIP Radford et al. (2021) and BLIP Li et al. (2022), have achieved remarkable progress on VQA. However, their performance often degrades under domain shifts Liu et al. (2025); Luo et al. (2019), where visual distributions and linguistic expressions differ from those seen during pretraining. As a result, VLMs trained on source domains exhibit limited generalization to specialized target domains Singhal et al. (2023). Test-time domain adaptation (TTDA) has emerged as an effective solution for improving model robustness under domain shifts Liang et al. (2025); Wang et al. (2025; 2022). Unlike conventional domain adaptation methods Yang & Soatto (2020); Xu et al. (2021), TTDA adapts models using only unlabeled target-domain data during inference. Existing VQA-oriented TTDA methods Wen et al. (2023); Liu et al. (2024); Farina et al. (2024) mainly mitigate language bias through self-supervised adaptation strategies, while general TTDA methods have demonstrated strong performance in image classification, segmentation, and multimodal tasks using confidence regularization, consistency learning, or feature alignment Wang et al. (2020); Mutlu et al. (2023); Wang et al. (2023); Zhang et al. (2024b); Lafon et al. (2025); Ebrahimi et al. (2022). These successes motivate extending TTDA to cross-domain VQA.

However, directly applying existing TTA methods to VQA remains challenging. Most TTA approaches Ma (2024); Ma et al. (2022); Cui et al. (2025); Adachi et al. (2024) formulate adaptation over a shared output space, implicitly assuming that all samples share a common output space. This assumption is appropriate for conventional classification tasks but does not naturally hold for VQA, where the valid answer space is conditioned on each image-question pair. Optimizing confidence over a shared answer vocabulary therefore introduces optimization coupling across unrelated questions, making optimization vulnerable to dominant-answer bias and question-insensitive predictions under domain shifts. Figure 1 illustrates the difference between conventional shared-space adaptation and our candidate-conditioned formulation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

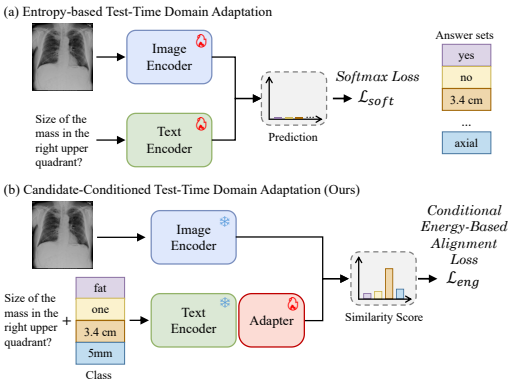


Figure 1: Conventional TTDA performs adaptation over a shared answer space, whereas our method performs adaptation within instance-specific candidate sets.

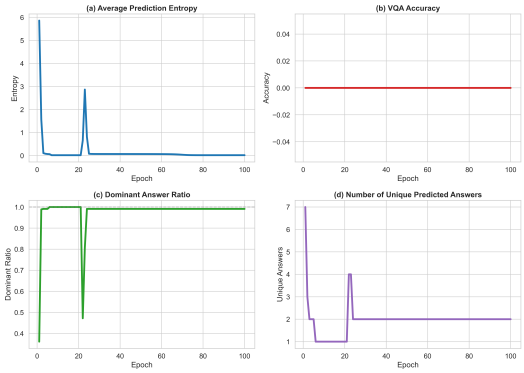


Figure 2: Entropy-only adaptation under the shared-answer setting. (a) Entropy decreases. (b) Accuracy remains near zero. (c) Dominant-answer ratio approaches 1.0. (d) The number of unique predicted answers rapidly collapses.

To address this problem, we propose a candidate-conditioned TTDA framework for VQA. Instead of adapting over a shared answer vocabulary, our method performs adaptation within instance-specific candidate sets using a local energy-based image-text alignment objective. This reformulation decouples optimization across unrelated samples while preserving question-conditioned reasoning. We further introduce a lightweight text adapter for efficient parameter updates while keeping the pretrained backbone frozen. Overall, our framework provides a stable adaptation objective tailored to the conditional prediction structure of VQA.

We evaluate our framework on six vision-language models and six cross-domain VQA benchmarks spanning medical imaging, pathology, art, and remote sensing. Extensive experiments demonstrate consistent improvements across diverse domains, achieving an average accuracy gain of 18.8% over direct candidate-based inference. Additional theoretical analysis, implementation details, ablation studies, and extended experimental results are provided in the **Appendix** to ensure reproducibility. The implementation code will be publicly released upon acceptance.

2 Failure of Shared-Space Entropy-Based TTDA in VQA

Entropy minimization Wang et al. (2020); Lee et al. (2024) is a widely used objective for TTDA. However, VQA requires predictions to remain conditioned on both images and questions, whereas entropy minimization encourages confident predictions over a shared answer space without explicitly preserving question-conditioned discrimination.

Problem Formulation. In VQA, predictions are conditioned on both an image x and a question q. Let the model parameterized by θ output: $p_\theta(y \mid x, q)$, where y belongs to a finite answer vocabulary $\mathcal{Y}$. Under unlabeled target-domain data, entropy-based TTDA minimizes the empirical entropy:

$$\mathcal{L}_{ent}(\theta) = \mathbb{E}_{(x,q)\sim\mathcal{D}_t}\left[-\sum_{y\in\mathcal{Y}} p_\theta(y \mid x, q)\log p_\theta(y \mid x, q)\right].$$

Unlike classification with a fixed label space, VQA is inherently question-dependent. Entropy minimization alone does not preserve question-conditioned discrimination.

Theoretical Analysis. We analyze why global entropy minimization may admit low-diversity solutions under shared parameter optimization. For any individual input pair (x, q), the predictive entropy is minimized when the output distribution is one-hot:

$$H(p_\theta(\cdot \mid x, q)) = 0 \quad \Longleftrightarrow \quad p_\theta(y^* \mid x, q) = 1 \text{ for some } y^*.$$

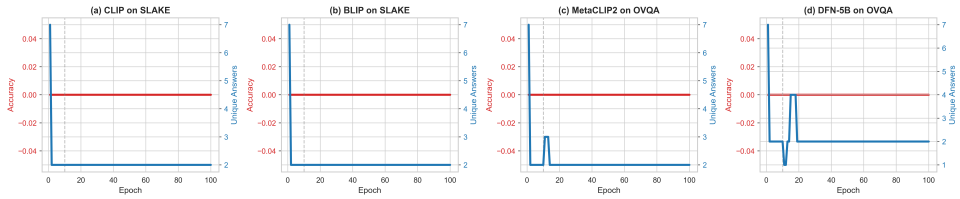


Figure 3: Structural collapse under entropy minimization across multiple VQA models and datasets.

In test-time adaptation, entropy is minimized over the entire target dataset:

$$\mathcal{L}_{ent}(\theta) = \frac{1}{N}\sum_{i=1}^{N} H\big(p_\theta(\cdot \mid x_i, q_i)\big).$$

Although each sample is minimized independently, the shared parameters couple optimization across all (x_i, q_i). Consequently, one feasible low-entropy solution under shared parameter optimization is to increase confidence toward the same dominant answer across samples:

$$p_\theta(y^* \mid x, q) = 1 \quad \forall (x, q) \in \mathcal{D}_t,$$

for a fixed answer $y^* \in \mathcal{Y}$. Such a constant predictor achieves zero entropy for every sample and therefore yields: $\mathcal{L}_{ent}(\theta) = 0$. Hence, constant predictors constitute one class of valid global minimizers of the entropy objective under shared-parameter optimization. In VQA models, predictions are computed as:

$$z = f_\theta(x, q), \quad p_\theta(y \mid x, q) = \text{softmax}(Wz + b).$$

If one answer initially receives slightly higher confidence because of initialization bias or answer imbalance, entropy minimization tends to further amplify its dominance. As the corresponding logit continues to dominate, the sensitivity of the prediction to the question embedding decreases:

$$\left\| \frac{\partial p_\theta(y \mid x, q)}{\partial e_q} \right\| \to 0.$$

The prediction therefore becomes increasingly less sensitive to the question embedding:

$$p_\theta(y|x, q) \approx p_\theta(y|x),$$

leading to a low-diversity solution that we refer to as structural collapse. This behavior arises from entropy-driven optimization under shared parameters rather than the model itself Sheng et al. (2025).

Empirical Evidence of Structural Collapse. Figure 2 illustrates entropy-based adaptation on VQA-RAD Lau et al. (2018) with CLIP. Although entropy decreases rapidly, accuracy remains near zero, while predictions collapse to a few dominant answers, as reflected by the increasing dominant-answer ratio and the decreasing number of unique predictions. Consistent with the theoretical analysis, entropy-based adaptation therefore converges to low-diversity solutions under shared parameter optimization. Figure 3 further shows that this phenomenon is not limited to CLIP or VQA-RAD. Across CLIP, BLIP, MetaCLIP2 Chuang et al. (2025), and DFN-5B Fang et al. (2023) on SLAKE Liu et al. (2021) and OVQA Huang et al. (2022), entropy-only adaptation collapses to nearly constant predictions with zero accuracy, indicating that structural collapse is a general failure mode.

Why Global Entropy Minimization Fails for VQA. The collapse stems from three factors: (1) entropy minimization promotes confident predictions without explicitly preserving conditional diversity; (2) VQA requires question-conditioned reasoning; and (3) answer imbalance amplifies dominant-answer bias. Formally, entropy-based TTA minimizes:

$$\min_\theta \mathbb{E}_{(x,q)\sim\mathcal{D}_t} H(Y|x, q),$$

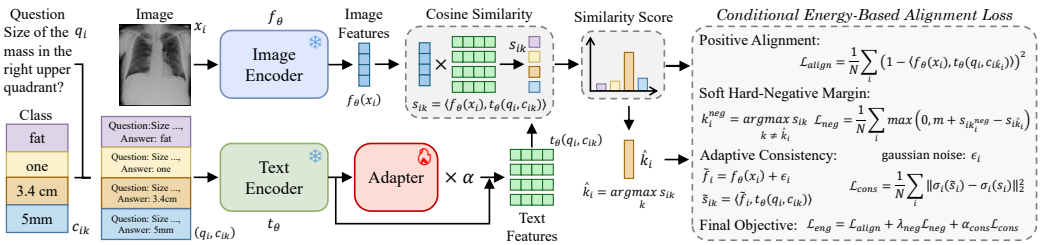


Figure 4: Overview of the proposed candidate-conditioned TTDA framework for VQA.

without preserving question sensitivity or conditional answer structure. Consequently, low-diversity solutions, including constant predictors, remain valid global minimizers.

3 Candidate-Conditioned Test-Time Domain Adaptation

Following the above analysis, we reformulate VQA test-time adaptation as a candidate-conditioned optimization problem, where adaptation is performed within an instance-specific candidate set using a local energy-based alignment objective instead of a shared answer vocabulary.

3.1 Conditional Candidate Optimization Space

Conventional entropy-based TTA assumes a shared global label space $\mathcal{Y}$ and optimizes:

$$\min_{\theta} \mathbb{E}_{(x,q)\sim\mathcal{D}_t} H\big(p_\theta(y \mid x, q)\big),$$

where $H(\cdot)$ denotes entropy. To eliminate this degeneracy, we perform adaptation within an instance-specific candidate set. For each test sample (x_i, q_i), we construct a *sample-specific candidate set*: $C_i = \{c_{i1}, \ldots, c_{ik}\}$.

Instead of predicting over a global vocabulary, the model performs conditional scoring only within C_i: $s_{ik} = \langle f_\theta(x_i), t_\theta(q_i, c_{ik})\rangle$, where f_θ and t_θ denote pretrained image and text encoders (optionally adapted via lightweight modules during TTDA). We then define the predicted index: $\hat{k}_i = \arg\max_k s_{ik}$, which serves as a pseudo-positive within the candidate set. Adaptation is therefore performed within each candidate set instead of a shared answer space.

3.2 Candidate Set Construction

We evaluate on six cross-domain VQA benchmarks, including VQA-RAD Lau et al. (2018), SLAKE Liu et al. (2021), PathVQA He et al. (2020), OVQA Huang et al. (2022), SemArt Garcia et al. (2020), and EarthVQA Wang et al. (2024). These datasets span diverse domains, including medical imaging, pathology, art understanding, and remote sensing, introducing substantial domain variation.

Closed-Answer Tasks. For datasets containing closed-answer questions (VQA-RAD, SLAKE, PathVQA, OVQA), samples labeled as Closed are evaluated using a canonical binary candidate: $C_i = \{\text{yes}, \text{no}\}$.

Open-Answer Tasks. For open-answer benchmarks, we follow the standard multiple-choice evaluation protocol adopted by prior candidate-based VQA benchmarks. Each candidate set contains one reference answer from the benchmark annotation and $K-1$ randomly sampled negatives from the predefined answer pool: $C_i = \{y_i^{gt}, c_{i1}, c_{i2}, \ldots, c_{i(K-1)}\}$, where the negative candidates are randomly sampled from $\mathcal{A}_{open}$ without replacement. We set $K = 4$ in the main experiments (one positive and three negatives). During adaptation, only the candidate set is available. Datasets without explicit answer-type annotations (SemArt and EarthVQA) are treated as fully open-answer tasks. This ensures that both closed- and open-answer tasks are evaluated under a unified candidate-based formulation.

Conditional Energy-Based Alignment Loss. Given the candidate set, we optimize a local energy-based alignment objective (Figure 4). The similarity score is:

$$s_{ik} = \langle f_\theta(x_i), t_\theta(q_i, c_{ik}) \rangle .$$

We use ℓ_2-normalized image and text features: $\|f_\theta(x)\|_2 = 1, \|t_\theta(q, c)\|_2 = 1$. We define the energy associated with candidate c_{ik} as: $E_{ik} = -s_{ik}$.

Positive Alignment. For each sample, we select the highest-scoring candidate as a pseudo-positive: $\hat{k}_i = \arg\max_k s_{ik}$. The alignment loss is:

$$\mathcal{L}_{align} = \frac{1}{N} \sum_i \left(1 - \left\langle f_\theta(x_i), t_\theta(q_i, c_{i\hat{k}_i}) \right\rangle\right)^2 .$$

Soft Hard-Negative Margin. To ensure clear separation within each candidate set, we identify the hardest negative: $k_i^{neg} = \arg\max_{k \neq \hat{k}_i} s_{ik}$.

We then impose a margin-based ranking constraint:

$$\mathcal{L}_{neg} = \frac{1}{N} \sum_i \max\big(0,\, m + s_{ik_i^{neg}} - s_{i\hat{k}_i}\big),$$

where $m > 0$ (default: 0.5) is a margin hyperparameter. This loss enforces the inequality: $s_{i\hat{k}_i} \geq s_{ik_i^{neg}} + m$.

Adaptive Consistency. To improve robustness to noisy pseudo labels, we introduce a feature-space consistency regularizer. Specifically, we perturb the image feature with Gaussian noise: $\tilde{f}_i = f_\theta(x_i) + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2 I)$ is isotropic Gaussian noise with small variance σ^2. After ℓ_2 normalization of the perturbed feature, we compute the image-text similarity scores:

$$\tilde{s}_{ik} = \langle \tilde{f}_i, t_\theta(q_i, c_{ik}) \rangle.$$

Let:

$$s_i = [s_{i1}, s_{i2}, \ldots, s_{ik}], \qquad \tilde{s}_i = [\tilde{s}_{i1}, \tilde{s}_{i2}, \ldots, \tilde{s}_{ik}],$$

denote the score vectors over the candidate set C_i. Let $\sigma_i(\cdot)$ denote the softmax operation applied within the candidate set C_i. We enforce prediction consistency between the original and perturbed representations:

$$\mathcal{L}_{cons} = \frac{1}{N} \sum_i \|\sigma_i(\tilde{s}_i) - \sigma_i(s_i)\|_2^2 .$$

This regularizer encourages locally stable predictions under feature perturbations.

Final Objective. The overall TTA objective is defined as:

$$\mathcal{L}_{eng} = \mathcal{L}_{align} + \lambda_{neg} \mathcal{L}_{neg} + \alpha_{cons} \mathcal{L}_{cons}.$$

In all experiments, we use default $\lambda_{neg} = 0.01$, $\alpha_{cons} = 0.1$.

3.3 Lightweight TTA Adapter Design

We freeze both encoders and insert a lightweight MLP adapter into the text branch. Given a normalized text feature $t \in \mathbb{R}^d$, we apply a gated residual modulation:

$$t' = \text{Normalize}\big(t + \alpha g_\phi(t)\big),$$

where $g_\phi(\cdot)$ is a two-layer multilayer perceptron (MLP) with ReLU activation. The scalar gate α controls the adaptation magnitude and is initialized to 0.5 by default. Only the adapter parameters ϕ and α are updated during adaptation, while all backbone parameters remain frozen. The adapter introduces a negligible number of additional parameters.

Table 1: Statistics of the datasets used for evaluating test-time domain adaptation.

Dataset		VQA-RAD	SLAKE	PathVQA	OVQA	SemArt	EarthVQA
Test data	images	315	96	1,000	1,223	21,384	1,809
	questions	451	1,061	6,012	1,902	4,912	63,225
	answer sets	113	130	1,300	275	3,189	112

Table 2: Entropy TTDA fails under full-answer-space VQA.

Model	VQA-RAD	SLAKE	OVQA
CLIP (full fine-tuning)	0.0	0.0	0.0
CLIP (LoRA)	0.0	0.0	0.0
CLIP (text adapter)	0.0	0.0	0.05
MetaCLIP2 (full fine-tuning)	0.0	0.0	0.0
MetaCLIP2 (LoRA)	0.0	0.0	0.0
MetaCLIP2 (text adapter)	0.0	0.0	0.03
DFN-5B (full fine-tuning)	0.0	0.0	0.0
DFN-5B (LoRA)	0.0	0.05	0.0
DFN-5B (text adapter)	0.0	0.0	0.0

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate our method on six cross-domain VQA benchmarks: VQA-RAD, SLAKE, PathVQA, OVQA, SemArt, and EarthVQA. Following the standard TTDA protocol, adaptation is performed on unlabeled target-domain test sets. Dataset statistics are summarized in Table 1.

Vision-Language Models. We evaluate six representative VLMs: CLIP Radford et al. (2021), BLIP Li et al. (2022), SigLIP2 Tschannen et al. (2025), MetaCLIP2 Chuang et al. (2025), DFN-2B Fang et al. (2023), and DFN-5B Fang et al. (2023). All models perform prediction by ranking candidate answers according to image-text similarity scores.

Implementation Details. We insert a lightweight text adapter while freezing the pretrained backbone Houlsby et al. (2019); Sung et al. (2022). Optimization uses AdamW Loshchilov & Hutter (2017) with learning rate 1×10^{-4} and weight decay 1×10^{-4}. All experiments are conducted on a single NVIDIA V100 GPU.

Evaluation Metrics. We report answer accuracy following the official protocols of each dataset. For VQA-RAD, SLAKE, PathVQA, and OVQA, Open, Closed, and Overall accuracies are reported. For SemArt and EarthVQA, Overall accuracy is reported. Details are provided in the Appendix.

4.2 FAILURE OF TENT TTA ON FULL-VOCABULARY VQA

Table 2 evaluates entropy-based adaptation under the conventional full-vocabulary VQA setting using full fine-tuning, LoRA, and our text adapter under the same entropy objective. All adaptation strategies produce near-zero accuracy across datasets. Combined with the collapse analysis in Figure 2, this indicates that the failure arises from entropy optimization over a shared answer vocabulary rather than from the choice of adaptation parameterization. Entropy minimization amplifies dominant answers, leading to low-diversity predictions that ignore question-dependent reasoning.

4.3 MAIN RESULTS: CANDIDATE-CONDITIONED TTDA

Table 3 summarizes the results on six VLMs and six VQA benchmarks. Compared with direct inference (Base), the proposed candidate-conditioned TTA consistently improves performance across all models and datasets, achieving an average gain of **18.8%**. These results demonstrate that optimizing within instance-specific candidate sets provides an effective and stable formulation for test-time adaptation in VQA. Figure 5 further illustrates the adaptation trend on VQA-RAD, where performance improves from 42.1% to 53.7% while maintaining stable convergence. Performance is also highly consistent across three random seeds (Table 4), confirming the robustness of the proposed adaptation. Among all benchmarks, EarthVQA exhibits the largest improvement. For example,

Table 3: Main results on six VQA datasets (accuracy, %). Base denotes direct inference, while Ours applies the proposed TTDA.

Models	Medical Radiology and Pathology												Art	Remote
	VQA-RAD			SLAKE			PathVQA			OVQA			SemArt	EarthVQA
	Open	Closed	Overall	Open	Closed	Overall	Open	Closed	Overall	Open	Closed	Overall	Overall	Overall
CLIP (Base, 151.3M)	33.9	47.6	42.1	37.2	47.4	41.2	32.7	57.2	45.0	31.7	62.7	50.4	36.4	0.0588
CLIP (Ours, 151.6M)	53.1	54.1	53.7	59.3	54.6	57.4	44.9	**59.6**	52.3	38.9	73.3	59.6	50.4	**91.1**
BLIP (Base, 361.2M)	23.0	47.5	37.7	18.3	48.1	30.0	16.0	47.4	31.7	15.6	57.9	41.0	19.6	0.116
BLIP (Ours, 361.8M)	50.2	48.8	49.4	47.2	52.7	49.3	49.9	50.9	50.4	35.4	72.5	57.7	40.8	25.4
SigLIP2 (Base, 375.2M)	40.8	52.6	47.9	34.5	54.4	42.3	31.0	46.6	38.8	36.2	49.7	44.3	34.7	0.0768
SigLIP2 (Ours, 375.6M)	51.9	53.6	52.9	47.3	54.9	50.4	43.2	50.3	46.7	32.5	72.6	56.6	44.1	27.3
MetaCLIP2 (Base, 1082.1M)	36.3	50.0	44.6	50.2	42.8	47.3	44.7	45.1	44.9	34.6	38.3	36.8	38.3	0.0804
MetaCLIP2 (Ours, 1082.5M)	**54.8**	**54.5**	**54.6**	55.3	54.5	55.0	**62.9**	50.4	**56.6**	51.2	61.8	57.6	56.1	51.6
DFN-2B (Base, 149.6M)	34.1	51.9	44.8	41.4	45.0	42.8	28.1	44.7	36.4	33.3	32.0	32.5	36.5	0.0528
DFN-2B (Ours, 150.1M)	51.9	54.0	53.2	56.3	54.6	55.6	56.1	49.0	52.5	55.7	41.3	47.1	54.1	35.6
DFN-5B (Base, 986.1M)	30.8	49.3	42.0	36.8	53.9	43.5	27.4	49.9	38.7	33.3	49.4	43.0	37.1	0.0425
DFN-5B (Ours, 988.1M)	52.0	53.3	52.8	**69.3**	**55.5**	**61.9**	58.5	52.3	55.4	**57.6**	**75.8**	**66.5**	**56.2**	36.5

Table 4: Robustness across three random seeds (mean % $\pm$ std).

Model	OVQA	SemArt	EarthVQA
CLIP	59.6 ± 0.2	50.4 ± 0.2	91.1 ± 0.2
MetaCLIP2	57.6 ± 0.2	56.1 ± 0.3	51.6 ± 0.4
DFN-5B	66.5 ± 0.4	56.2 ± 0.2	36.5 ± 0.2

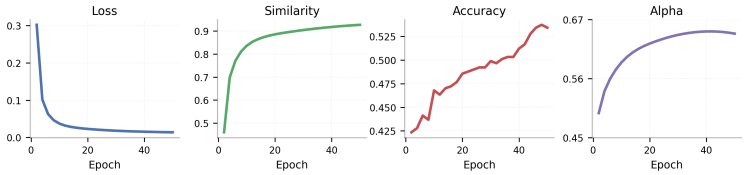


Figure 5: Adaptation dynamics on VQA-RAD (CLIP).

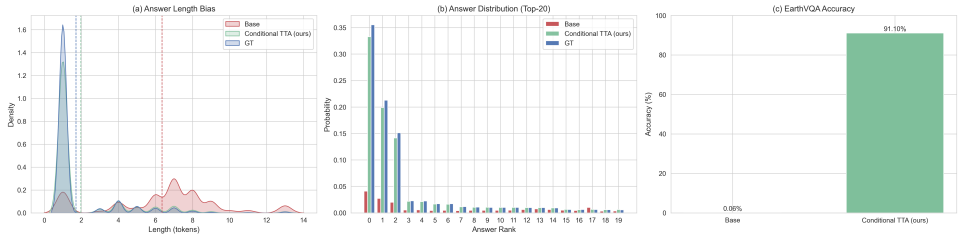


Figure 6: Diagnostic analysis of CLIP on EarthVQA. (a) Predicted answer lengths. (b) Prediction frequency over 112 candidate answers. (c) Overall accuracy.

CLIP increases from **0.0588%** to **91.1%**. Given this unusually large improvement, we further analyze this phenomenon in the next subsection and the Appendix. The extremely low baseline is caused by a structural mismatch between pretrained image-text similarity and the discrete EarthVQA answer space rather than implementation errors or optimization randomness.

4.4 Understanding the EarthVQA Performance Gain

Figure 6 analyzes the unusually low CLIP baseline (0.0588%) on EarthVQA. Both the base model and our method use the same four-way candidate-conditioned inference protocol; the only difference is whether conditional alignment is performed during test-time domain adaptation. Figure 6(a) shows that the pretrained model consistently favors longer answers than the ground-truth distribution, indicating a systematic bias in the pretrained similarity space. Conditional alignment substantially reduces this bias. Figure 6(b) further reveals severe answer distribution collapse, where predictions concentrate on a small subset of generic candidates. In contrast, our method produces a much more balanced distribution by improving similarity ranking within the candidate space. Figure 6(c) confirms that the near-zero baseline arises from a structural mismatch between pretrained similarity geometry and the discrete EarthVQA answer space rather than implementation issues or insufficient model capacity. Small ranking biases accumulate into systematic prediction errors

Table 5: Comparison between entropy-based TTA (Tent) and our alignment loss.

Model	Loss	VQA-RAD	SLAKE	OVQA
CLIP	No Adaptation	42.1	41.2	50.4
	Entropy (Tent)	42.8	33.9	48.2
	Alignment (Ours)	**53.7**	**57.4**	**59.6**
MetaCLIP2	No Adaptation	44.6	47.3	36.8
	Entropy (Tent)	42.6	48.9	54.4
	Alignment (Ours)	**54.6**	**55.0**	**57.6**
DFN-5B	No Adaptation	42.0	43.5	43.0
	Entropy (Tent)	42.6	31.2	55.1
	Alignment (Ours)	**52.8**	**61.9**	**66.5**

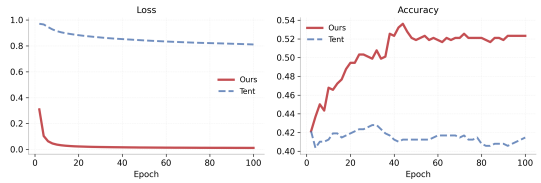


Figure 7: Optimization dynamics of entropy minimization and the proposed alignment objective.

across the benchmark. By optimizing conditional alignment under the same inference protocol, our method adapts the similarity space to the target answer distribution while preserving the original retrieval mechanism, improving CLIP from 0.0588% to 91.1%. A similar prediction bias is also observed in the other five VLMs, all of which consistently benefit from the proposed adaptation.

4.5 Entropy Minimization vs Conditional Alignment

We compare Tent Wang et al. (2020), which minimizes prediction entropy, against our conditional alignment objective under the same candidate answer space. As shown in Table 5, entropy minimization yields limited and inconsistent improvements and even degrades performance in several cases. In contrast, our objective improves performance across all models and datasets. Figure 7 further shows that our method converges rapidly with stable accuracy gains, whereas Tent reduces prediction entropy but fails to improve accuracy.

4.6 Comparison with Other TTA Methods

We compare our method with representative TTA methods, including Tent Wang et al. (2020), TPT Zhu et al. (2024), VPA Sun et al. (2023), ZERO Farina et al. (2024), BoostAdapter Zhang et al. (2024b), and DocTTA Ebrahimi et al. (2022), using CLIP under the same candidate-conditioned evaluation protocol. As shown in Table 6, existing methods improve performance to varying degrees, whereas our method consistently achieves the best performance across all datasets.

4.7 Comparison with Specialized VQA-TTA Methods

We compare the proposed method with two representative VQA-specific test-time adaptation methods, TDS Wen et al. (2023) and QED Liu et al. (2024). Unlike existing VQA-specific TTA methods, which rely on self-supervised adaptation or question-type-aware debiasing, our approach explicitly performs candidate-conditioned alignment during adaptation. All methods are evaluated under the same candidate-conditioned evaluation protocol. Table 7 reports results under both offline multiple-step adaptation and the more challenging online single-step setting. Our method consistently achieves the best performance across models and datasets in both settings. Even under strict online adaptation, where each test sample is observed only once, our method remains consistently superior or at least competitive.

4.8 Ablation Study on Adaptation Components

Table 8 compares different adaptation strategies. Adapting only the text encoder with the proposed adapter consistently achieves the best performance across all models and datasets, indicating that

Table 6: Comparison of TTDA methods on CLIP under the same candidate-conditioned setting.

Dataset	No Adapt.	Tent	TPT	VPA	ZERO	Boost	DocTTA	Ours
VQA-RAD	42.1	42.8	46.9	44.6	43.5	45.2	46.0	**53.7**
SLAKE	41.2	33.9	48.1	38.5	42.6	42.4	45.5	**57.4**
OVQA	50.4	48.2	54.2	50.6	51.8	51.5	52.8	**59.6**

Table 7: Comparison with existing VQA-TTA methods under offline and online adaptation.

Model	Method	VQA-RAD		SLAKE		OVQA		EarthVQA	
		multiple	single	multiple	single	multiple	single	multiple	single
CLIP	-	42.1	42.1	41.2	41.2	50.4	50.4	0.0588	0.0588
	TDS	48.2	41.7	49.6	42.6	55.4	52.0	46.8	4.2
	QED	49.6	**42.5**	51.0	44.4	57.2	52.6	51.6	4.7
	Ours	**53.7**	**42.5**	**57.4**	**45.0**	**59.6**	**54.4**	**91.1**	**10.6**
MetaCLIP2	-	44.6	44.6	47.3	47.3	36.8	36.8	0.0804	0.0804
	TDS	49.9	45.2	51.3	49.6	49.4	48.2	31.6	3.8
	QED	51.1	45.6	53.2	50.3	50.9	48.6	36.7	4.1
	Ours	**54.6**	**46.2**	**55.0**	**52.5**	**57.6**	**48.9**	**51.6**	**8.7**
DFN-5B	-	42.0	42.0	43.5	43.5	43.0	43.0	0.0425	0.0425
	TDS	47.6	46.4	50.6	45.3	49.6	47.9	16.8	2.3
	QED	49.1	46.8	53.4	46.1	51.2	49.8	18.2	2.8
	Ours	**52.8**	**47.4**	**61.9**	**48.7**	**66.5**	**63.5**	**36.5**	**5.6**

Table 8: Ablation study on different adaptation components and parameterization strategies.

Models	Methods	VQA-RAD	SLAKE	OVQA
CLIP	No Adaptation	42.1	41.2	50.4
	Image encoder + Adapter	44.8	42.7	53.6
	Text encoder + Adapter	**53.7**	**57.4**	**59.6**
	Image-Text encoder + Adapter	45.1	46.7	52.9
	Text encoder + LoRA	44.4	52.0	55.1
	Image-Text encoder + LoRA	43.1	53.4	57.3
	All Adaptation	43.6	45.0	53.1
MetaCLIP2	No Adaptation	44.6	47.3	36.8
	Image encoder + Adapter	44.9	47.5	37.4
	Text encoder + Adapter	**54.6**	**55.0**	**57.6**
	Image-Text encoder + Adapter	45.7	49.3	38.6
	Text encoder + LoRA	45.3	49.5	40.6
	Image-Text encoder + LoRA	45.1	39.3	38.2
	All Adaptation	44.6	48.5	50.6
DFN-5B	No Adaptation	42.0	43.5	43.0
	Image encoder + Adapter	45.1	48.0	49.0
	Text encoder + Adapter	**52.8**	**61.9**	**66.5**
	Image-Text encoder + Adapter	46.4	45.6	64.1
	Text encoder + LoRA	48.3	50.6	66.3
	Image-Text encoder + LoRA	46.4	44.0	62.7
	All Adaptation	42.6	43.8	62.0

Table 9: Ablation of the proposed objective components.

Model	Alignment	Negative	Consistency	VQA-RAD	SLAKE	OVQA
CLIP	✓			52.3	52.7	54.3
	✓	✓		52.7	54.2	56.4
	✓		✓	53.1	55.0	56.6
	✓	✓	✓	**53.7**	**57.4**	**59.6**
MetaCLIP2	✓			53.4	51.6	56.4
	✓	✓		54.0	53.4	56.9
	✓		✓	53.7	53.0	57.3
	✓	✓	✓	**54.6**	**55.0**	**57.6**
DFN-5B	✓			52.1	54.6	58.2
	✓	✓		52.3	57.6	60.4
	✓		✓	52.5	58.0	61.2
	✓	✓	✓	**52.8**	**61.9**	**66.5**

text-side adaptation is most effective for candidate-conditioned VQA. Adapting only the image encoder provides limited gains, while updating both encoders or all parameters often reduces stability, suggesting that excessive parameter updates disrupt the pretrained vision-language alignment. Compared with LoRA, our lightweight adapter also delivers consistently better performance, validating its effectiveness for efficient test-time adaptation.

4.9 Ablation Study of the Proposed Objective Loss

Table 9 reports the contribution of each objective component. Starting from the alignment objective alone, both the hard-negative margin and regularization improve performance across all models and datasets. The complete objective achieves the best results in every setting, demonstrating that the three components provide complementary benefits for robust TTDA.

5 Conclusion

In this paper, we identify a structural mismatch between conventional TTDA objectives and VQA. Entropy-based TTDA may fail in this setting because minimizing prediction entropy over a shared answer space admits degenerate constant predictors, leading to structural collapse where predictions become insensitive to the input question. To address this issue, we propose a candidate-conditioned test-time adaptation framework that restricts predictions to instance-specific candidate sets and aligns image and text features, enabling stable test-time adaptation with a lightweight text adapter. Extensive experiments across multiple VLMs and VQA benchmarks demonstrate consistent improvements under domain shifts, highlighting the importance of preserving question-conditioned reasoning during test-time adaptation rather than relying solely on global entropy minimization.

AI USE STATEMENT

Generative AI tools were used for language editing and polishing, LaTeX formatting, table and figure layout adjustments, and limited assistance in editing plotting and formatting code. All AI-assisted outputs were reviewed and verified by the authors. Generative AI tools were not used to generate research ideas, develop the proposed method, design or conduct experiments, analyze experimental results, or formulate scientific claims. The authors take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

The paper provides the method formulation and optimization objectives, describes the datasets, vision-language models, implementation details, evaluation metrics, and hyperparameter settings used in the experiments, and includes further theoretical analysis, experimental settings, ablations, and diagnostic results in the Appendix.

REFERENCES

Kazuki Adachi, Shin'ya Yamaguchi, Atsutoshi Kumagai, and Tomoki Hamagami. Test-time adaptation for regression by subspace alignment. *arXiv preprint arXiv:2410.03263*, 2024.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, et al. Meta clip 2: A worldwide scaling recipe. *arXiv preprint arXiv:2507.22062*, 2025.

Shuang Cui, Jinglin Xu, Yi Li, Xiongxin Tang, Jiangmeng Li, Jiahuan Zhou, Fanjiang Xu, Fuchun Sun, and Hui Xiong. Bayestta: Continual-temporal test-time adaptation for vision-language models via gaussian discriminant analysis. *arXiv preprint arXiv:2507.08607*, 2025.

Mario Döbler, Robert A Marsden, Tobias Raichle, and Bin Yang. A lost opportunity for vision-language models: a comparative study of online test-time adaptation for vision-language models. In *European Conference on Computer Vision*, pp. 117–133. Springer, 2024.

Sayna Ebrahimi, Sercan O Arik, and Tomas Pfister. Test-time adaptation for visual document understanding. *arXiv preprint arXiv:2206.07240*, 2022.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023.

Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models. *Advances in Neural Information Processing Systems*, 37:129062–129093, 2024.

Noa Garcia, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Teruko Mitamura. A dataset and baselines for visual question answering on art. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pp. 92–108. Springer, 2020.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8340–8349, 2021.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Yefan Huang, Xiaoli Wang, Feiyan Liu, and Guofeng Huang. Ovqa: A clinically generated visual question answering dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2924–2938, 2022.

Dahuin Jung. Harnessing consistency for improved test-time adaptation. *Image and Vision Computing*, pp. 105650, 2025.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14162–14171, 2024.

Marc Lafon, Gustavo Adolfo Vargas Hakim, Clément Rambour, Christian Desrosier, and Nicolas Thome. Cliptta: Robust contrastive vision-language test-time adaptation. *arXiv preprint arXiv:2507.14312*, 2025.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. *arXiv preprint arXiv:2403.07366*, 2024.

Jingjing Li, Zhiqi Yu, Zhekai Du, Lei Zhu, and Heng Tao Shen. A comprehensive survey on source-free domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8): 5743–5762, 2024.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4654–4662, 2019.

Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1):31–64, 2025.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1650–1654. IEEE, 2021.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Jin Liu, Jialong Xie, Fengyu Zhou, and Shengfeng He. Question type-aware debiasing for test-time visual question answering model adaptation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11):10805–10816, 2024.

Lei Liu, Xiangdong Su, and Guanglai Gao. Fourier self-adaptation for transferring general pre-trained models to specific domains. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3605–3614, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2507–2516, 2019.

Jing Ma. Improved self-training for test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23701–23710, 2024.

Wenao Ma, Cheng Chen, Shuang Zheng, Jing Qin, Huimao Zhang, and Qi Dou. Test-time adaptation with calibration of medical image classification nets for label distribution shift. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 313–323. Springer, 2022.

Onur Cezmi Mutlu, Mohammadmahdi Honarmand, Saimourya Surabhi, and Dennis P Wall. Tempt: Temporal consistency for test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5917–5923, 2023.

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*, pp. 8748–8763. PMLR, 2021.

Lijun Sheng, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. The illusion of progress? a critical look at test-time adaptation for vision-language models. *arXiv preprint arXiv:2506.24000*, 2025.

Akshit Singh, Shyam Marjit, Wei Lin, Paul Gavrikov, Serena Yeung-Levy, Hilde Kuehne, Rogerio Feris, Sivan Doveh, James Glass, and M Jehanzeb Mirza. Ttrv: Test-time reinforcement learning for vision language models. *arXiv preprint arXiv:2510.06783*, 2025.

Peeyush Singhal, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. Domain adaptation: challenges, methods, datasets, and applications. *IEEE access*, 11:6973–7020, 2023.

Jiachen Sun, Mark Ibrahim, Melissa Hall, Ivan Evtimov, Z Morley Mao, Cristian Canton Ferrer, and Caner Hazirbas. Vpa: Fully test-time visual prompt adaptation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 5796–5806, 2023.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5227–5237, 2022.

Yushun Tang, Shuoshuo Chen, Zhehan Kan, Yi Zhang, Qinghai Guo, and Zhihai He. Learning visual conditioning tokens to correct domain shift for fully test-time adaptation. *IEEE Transactions on Multimedia*, 2024.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Jeya Maria Jose Valanarasu, Pengfei Guo, Vishal M Patel, et al. On-the-fly test-time adaptation for medical image segmentation. In *Medical Imaging with Deep Learning*, pp. 586–598. PMLR, 2024.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

Junjue Wang, Zhuo Zheng, Zihang Chen, Ailong Ma, and Yanfei Zhong. Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5481–5489, 2024.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7201–7211, 2022.

Shuai Wang, Daoan Zhang, Zipei Yan, Jianguo Zhang, and Rui Li. Feature alignment and uniformity for test time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20050–20060, 2023.

Zixin Wang, Yadan Luo, Liang Zheng, Zhuoxiao Chen, Sen Wang, and Zi Huang. In search of lost online test-time adaptation: A survey. *International Journal of Computer Vision*, 133(3): 1106–1139, 2025.

Zhiquan Wen, Shuaicheng Niu, Ge Li, Qingyao Wu, Mingkui Tan, and Qi Wu. Test-time model adaptation for visual question answering with debiased self-supervisions. *IEEE Transactions on Multimedia*, 26:2137–2147, 2023.

Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 69–77, 2016.

Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14383–14392, 2021.

Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4085–4095, 2020.

Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *Proceedings of the european conference on computer vision (ECCV)*, pp. 300–317, 2018.

Longhui Yuan, Shuang Li, Zhuo He, and Binhui Xie. Few clicks suffice: Active test-time adaptation for semantic segmentation. *arXiv preprint arXiv:2312.01835*, 2023.

Maxime Zanella, Clément Fuchs, Christophe De Vleeschouwer, and Ismail Ben Ayed. Realistic test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25103–25112, 2025.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024a.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems*, 35:38629–38642, 2022.

Siqi Zhang, Lu Zhang, and Zhiyong Liu. Test-time adaptation for object detection via dynamic dual teaching. *Image and Vision Computing*, pp. 105740, 2025.

Taolin Zhang, Jinpeng Wang, Hang Guo, Tao Dai, Bin Chen, and Shu-Tao Xia. Boostadapter: Improving vision-language test-time adaptation via regional bootstrapping. *Advances in Neural Information Processing Systems*, 37:67795–67825, 2024b.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

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.

Yuhan Zhu, Guozhen Zhang, Chen Xu, Haocheng Shen, Xiaoxin Chen, Gangshan Wu, and Limin Wang. Efficient test-time prompt tuning for vision-language models. *arXiv preprint arXiv:2408.05775*, 2024.