This CVPR Findings paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

Dual-Modality Anchor-Guided Filtering for Test-time Prompt Tuning

Jungwon Choi Eunwoo Kim*
School of Computer Science and Engineering, Chung-Ang University
{severe0713, eunwoo}@cau.ac.kr

Abstract

Test-Time Prompt Tuning (TPT) adapts vision-language models using augmented views, but its effectiveness is hindered by the challenge of determining which views are beneficial. Standard entropy-based filtering relies on the internal confidence scores of the model, which are often miscalibrated under distribution shift, assigning high confidence to irrelevant crops or background regions while ignoring semantic content. To address this, we propose a dual-modality anchor-guided framework that grounds view selection in semantic evidence. We introduce a text anchor from attribute-rich descriptions, to provide fine-grained class semantics, and an adaptive image anchor that captures evolving test-time statistics. Using these anchors, we filter views based on alignment and confidence, ensuring that only informative views guide adaptation. Moreover, we treat the anchors as auxiliary predictive heads and combine their predictions with the original output in a confidence-weighted ensemble, yielding a stable supervision signal for prompt updates. Extensive experiments on 15 benchmark datasets demonstrate new state-of-the-art performance, highlighting the contribution of anchor-guided supervision as a foundation for robust prompt updates.

1. Introduction

Vision-language models (VLMs) [12, 25, 32, 40] have achieved remarkable zero-shot performance by aligning images and text within a shared semantic space. It enabled a wide range of visual tasks guided by natural language prompts. Since the effectiveness of these models is highly dependent on the quality of the prompts [6, 24, 25, 42], recent studies have explored both hand-crafted textual templates [2, 12, 25] and learnable prompts [13, 28, 41, 42] to improve performance. However, both approaches tend to overfit the training data distribution [17, 36], making it difficult to generalize to unseen domains. Test-time prompt tuning (TPT) [30] addresses this limitation by updating prompts directly on unlabeled data during inference, allowing models to better generalize to unseen domains without requiring additional labeled samples. This setting is particularly practical, as labeled target samples are rarely available in real-world scenarios. For each test sample, TPT generates predictions from multiple augmented views and updates the learnable prompt by minimizing entropy of average prediction, thereby adapting the model to the current test sample.

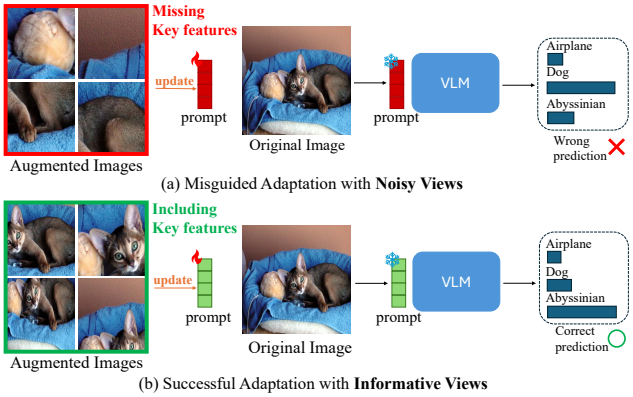

Figure 1. An example of how prompt quality varies depending on the selected views. When the (a) selected views are noisy and fail to capture key object features, the prompt is adapted in an incorrect direction, resulting in erroneous predictions. In contrast, (b) informative views containing essential features guide the prompt toward the correct semantics, resulting in accurate predictions. This highlights the critical role of robust view selection.

However, TPT methods are vulnerable to noisy or uninformative views generated by standard augmentations (e.g., random crop or resize). Many of these views miss key object regions, over-represent the background, or include content that weakly relates to the target class. Minimizing the entropy of such unreliable views is dangerous; it reinforces confidence of the model in wrong predictions and misguides the prompt adaptation toward erroneous semantics, as shown in Fig. 1. In contrast, well-selected and semantically aligned views play a crucial role; they correct the mistaken predictions of the original image and guide the prompt updates toward the true class direction. This ob-

*Corresponding author.

servation underscores that the effectiveness of TPT fundamentally depends on the quality and semantic fidelity of the selected views. Despite its importance, the problem of view selection has not been largely explored in prior TPT literature. Prior methods [1, 30, 37] rely solely on filtering views based on prediction entropy, selecting only those views that exhibit low prediction entropy. But these approaches are not sufficient, as they have two critical limitations. First, under distribution shift, prediction entropy of the model is often unreliable, leading to overconfident and inaccurate predictions [8, 15, 34]. This causes semantically uninformative views to be erroneously selected. Second, entropy is typically calculated using a coarse initial prompt (e.g.,"a photo of a [CLASS]"), which lacks the ability to distinguish fine-grained attributes [16, 38] necessary to verify its identity within a confusing crop.

One recent approach [8] extends entropy-based filtering by measuring both entropy and the cosine similarity to the original image. Although this proves effective in that setting, the improvement largely stems from the use of diffusion [27] based augmentations, which generate highly diverse views. In contrast, standard augmentations remain the practical choice for TPT methods [1, 30, 37, 39] due to their efficiency and broad applicability. However, they do not support this similarity-based criterion effectively. For example, similarity-based selection with standard augmentations reduces the diversity of views and weakens generalization, since it favors views visually similar to the original image. Notably, we find that once view selection is performed in a semantically-aware manner, even standard augmentations can surpass diffusion-based augmentations, while avoiding the substantial computational cost of diffusion (see Section 4.2).

To address the limitations of unreliable entropy and rigid visual similarity metrics, we propose a dual-modality anchor-guided framework for robust view selection. To form a fine-grained anchor and move beyond the limitations of coarse initial prompts, we leverage large language models (LLMs) to generate attribute-rich textual descriptions that capture detailed visual properties. Each description is weighted by its similarity to the image features and aggregated into a text anchor to better capture the semantic content of the image. We first use the text anchors to guide initial view filtering by measuring alignment and entropy. To further incorporate domain-specific visual cues and capture appearance variations beyond textual semantics, we introduce an image anchor that is updated via the cumulative average of previously selected views. We then employ the image anchor to guide the view selection based on alignment and entropy. By combining text and image anchors, our framework captures semantic fidelity and visual consistency, enabling reliable view validation even under distribution shift where entropy and direct similarity measures often fail. This approach is also lightweight and modular, enabling seamless integration into existing prompt tuning pipelines [13, 41, 42] with minimal additional cost.

Finally, we leverage the filtered set of views for the prompt adaptation. However, simply minimizing the entropy of these views, a common practice in prior works [8, 30, 37], can encourage biased updates. To address this, we further leverage our dual anchors not only as filters but also as auxiliary predictive sources. We present a confidence-aware ensemble that dynamically weights and integrates predictions from each source based on its maximum prediction probability. This forms a highly reliable target distribution, built from the robust consensus of the dual anchors and the original prompt. The prompts are then updated by minimizing a KL-divergence loss against this target distribution, providing stable and semantically grounded supervision for robust adaptation. The main contributions are:

- Our framework mitigates the limitations of entropy-based filtering by grounding the selection process in dual-modality anchors and jointly evaluating both alignment and confidence.
- We replace coarse prompts with dual-modality anchors, where the text anchor represents semantic textual information and the image anchor captures test distribution visual features.
- We leverage the anchors as predictive sources in a confidence-weighted ensemble, yielding stable supervision from augmented views.
- Extensive experiments across 15 diverse benchmarks demonstrate the effectiveness of our approach, achieving an average accuracy improvement of 3.36% over all compared methods.

2. Related Work

2.1. Prompt Learning

Prompt learning for vision-language models (VLMs) [12, 25] has been widely explored to enhance their performance on various downstream tasks. Inspired by prompt tuning in NLP [29], a common approach in vision tasks involves learning prompts using a small set of labeled examples [13, 41, 42]. CoOp [42] adapts VLMs by optimizing a set of continuous prompt vectors in its language branch. CoCoOp [41] extends CoOp [42] by conditioning prompts on individual image features. MaPLe [13] leverages multi-modal prompt learning with vision–language prompts. Co-Prompt [28] enforces a consistency constraint to prevent overfitting on downstream tasks. Any-Shift prompting [36] introduces adaptive prompts designed to handle arbitrary distribution shifts.

However, these prior works often rely on labeled samples from the target domain, which restricts their applicability in practical scenarios. In contrast, we tackle a more realis-

tic test-time prompt tuning setting, where the model must adapt on-the-fly using only unlabeled test data without access to target domain labels. This paradigm is particularly important for real-world deployment, enabling robust adaptation to novel and unseen environments where labeled data is unavailable.

2.2. Test-time Prompt Tuning

Test-time prompt tuning (TPT) [30] is an emerging paradigm that adapts vision-language models by optimizing prompts using unlabeled test samples during inference. TPT optimizes prompts by minimizing the entropy of the averaged predictions from augmented images. Building upon this, DiffTPT [8] incorporates a pre-trained diffusion model [27] to generate diverse augmentations. PromptAlign [1], building on MaPLe [13], explicitly aligns the mean and variance of image token embeddings from a proxy source dataset. C-TPT [39] improves test-time prompt tuning by maximizing text feature dispersion. DynaPrompt [37] introduces a prompt queue and a scoring mechanism that enables multiple prompt updates.

The aforementioned approaches suffer from key drawbacks: DiffTPT [8] is computationally heavy, PromptAlign [1] has limited applicability, and DynaPrompt [37] incurs extra memory and computation overhead. More critically, they all rely on entropy-based filtering, which is semantically blind and reinforces model biases [8, 43]. In contrast, our dual-modality, anchor-guided filtering approach resolves these issues by creating a robust, semantically grounded supervision signal, while serving as a lightweight and versatile module that can be easily integrated into various prompt tuning frameworks.

3. Method

Standard entropy-based confidence selection [1, 30, 37, 39], retains low-uncertainty views but often fails to exclude noisy or semantically irrelevant augmentations. To address this, we introduce a dual-modality anchor mechanism that guides view selection with semantic evidence as illustrated in Fig. 2. Our approach leverages text anchors derived from attribute-rich descriptions and adaptive image anchors accumulated from test-time embeddings, forming the basis for a more robust and informative selection process.

3.1. Text-Guided Semantic Filtering

To enhance the discriminative alignment between visual inputs and textual descriptions, we propose a filtering method that utilizes LLM-generated descriptions. We treat these descriptions as proxy attributes that are both *semantically rich* and *class-specific*. This approach of leveraging fine-grained semantic attributes over coarse, label-only prompts can be aligned with recent work in attribute-based recognition [19, 20, 23, 24], which has also highlighted their importance in enhancing discriminative power.

Attribute-Rich Description Generation. To enrich the textual representation, we leverage LLMs to generate attribute-oriented descriptions for each class. Following the practice in [23], we prompt an LLM with a diverse set of instructions, such as *"Describe what a(n) [CLASS] looks like"*. This process yields multiple attribute-rich descriptions per class, which are encoded by the text encoder into a knowledge base of textual attributes. This knowledge base provides stronger semantic grounding than conventional template prompts [12, 25] and serves as the foundation for our filtering process. All descriptions are generated and encoded offline as a one-time preprocessing step; during test time, our method only relies on lightweight similarity computations with pre-encoded text features, incurring negligible additional cost.

Alignment-Weighted Aggregation. LLM-generated descriptions provide diverse semantically rich attributes, but their relevance varies depending on the specific visual instance. To effectively prioritize descriptions aligned with the actual image content, we employ an adaptive aggregation mechanism. We are given image features from the augmented views, denoted as $\{\mathbf{e}_b\}_{b=1}^{B}$, where B is the number of views. Similarly, we have text description features $\{\mathbf{w}_c^i\}_{i=1}^{N}$, where N is the number of descriptions for each class c. All features are obtained from CLIP encoders [25]. For each tuple $\rho \triangleq (b, c, i)$, we compute an alignment score: $s_\rho = \mathrm{sim}(\mathbf{e}_b, \mathbf{w}_c^i) - \mathrm{sim}(\mathbf{e}_b, \bar{\mathbf{w}}_c)$, where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity and $\bar{\mathbf{w}}_c = \frac{1}{N}\sum_{i=1}^{N}\mathbf{w}_c^i$ is the class-wise mean feature. This score quantifies the *marginal benefit* of a specific description $\mathbf{w}_c^i$ over the generic, class-wise mean feature $\bar{\mathbf{w}}_c$. A high score indicates that the description provides semantic details that are particularly relevant to the image view.

To ensure a discriminative signal even under low-variance scenarios, we convert these scores into per-view weights $a_{b,c,i}$ via a softmax operation: $a_{b,c,i} = \frac{\exp(s_{b,c,i})}{\sum_{j=1}^{N}\exp(s_{b,c,j})}$. These weights represent the relevance of description i (for class c) to a specific view b. We then average these weights across all B views to obtain the final class-level description weights $\bar{a}_{c,i} = \frac{1}{B}\sum_{b=1}^{B} a_{b,c,i}$. Finally, the text anchor $\mathbf{t}_c$ is formed by an alignment-weighted aggregation: $\mathbf{t}_c = \sum_{i=1}^{N}\bar{a}_{c,i}\mathbf{w}_c^i$. Note that this entire process is performed adaptively for every single test image.

Joint Alignment-Confidence Scoring. We define a composite score for each view by augmenting an entropy-based uncertainty measure with a semantic alignment score. For a given view feature $\mathbf{e}$ and the number of class C, we compute

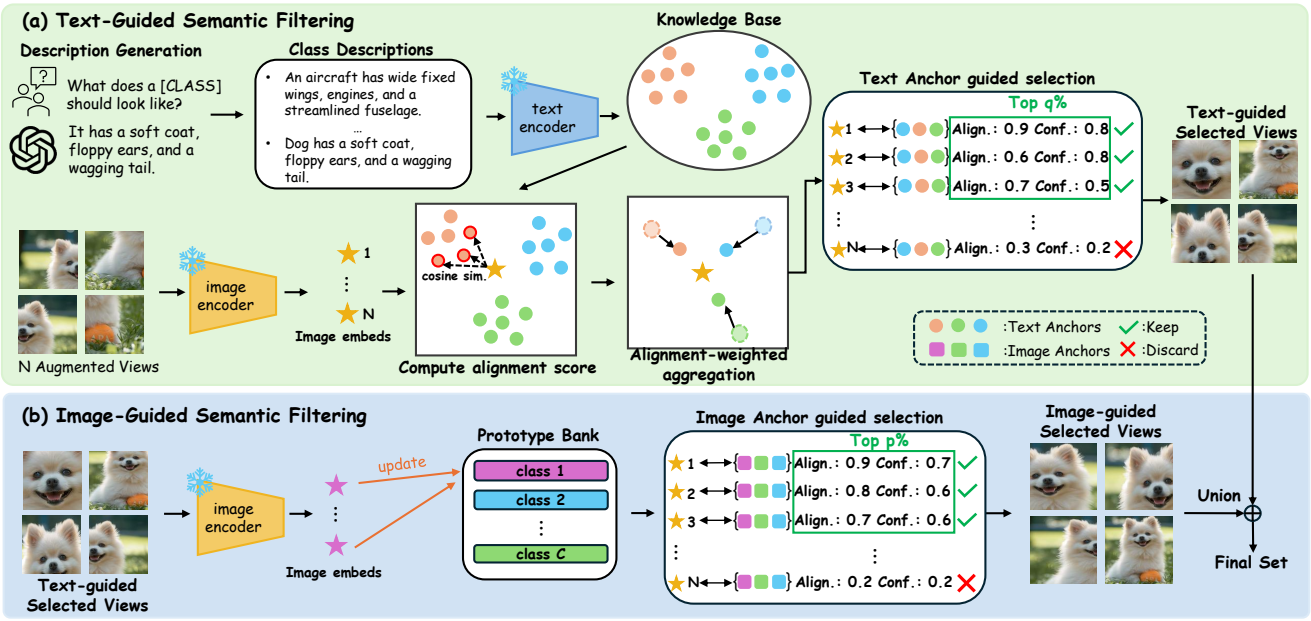

Figure 2. Overview of the proposed dual-modality anchor framework. The process begins with (a) text anchor construction. Multiple LLM-generated descriptions per class are encoded with the CLIP text encoder, and N augmented images are encoded with the CLIP image encoder. We then compute image–text alignment by measuring cosine similarity and perform alignment-weighted aggregation to form text anchors. Using these anchors, we compute joint alignment–confidence scores to each views and keep the top q% of N augmented views. (b) In the image anchor construction stage, the text-guided selected views are aggregated into a prototype bank that captures the visual distribution. The image anchor set is obtained by collecting the top-K class prototypes. We again compute joint alignment–confidence scores with the image anchors and retain the top p% of N augmented views. The final view set is the union of the two filters.

two components: alignment and confidence.

$$s_{\text{align}}(\mathbf{e}) = \max_{c \in \mathcal{C}} \text{sim}(\mathbf{e}, \mathbf{t}_c), \quad s_{\text{conf}}(\mathbf{e}) = 1 - \frac{\text{H}(p(y \mid \mathbf{e}))}{\log C}. \tag{1}$$

$\text{H}(\cdot)$ denotes the prediction entropy, and the normalized entropy $\frac{\text{H}(p(y|\mathbf{e}))}{\log C} \in [0,1]$ measures the uncertainty of the per-view predictive distribution: $p(y \mid \mathbf{e}) = \text{Softmax}(\tau \cdot \mathbf{e}^\top \mathbf{t}_c)$ where y is class label, τ is a learned temperature parameter, and the probabilities are computed from the similarity between the image feature and text anchors. The final score is a weighted sum of alignment and confidence:

$$\text{S}_{\text{text}}(\mathbf{e}) = \alpha_1 s_{\text{align}}(\mathbf{e}) + \alpha_2 s_{\text{conf}}(\mathbf{e}), \tag{2}$$

where the coefficients (α_1, α_2) balance the trade-off between semantic alignment and uncertainty. We retain the top-$q\%$ of views with the highest S_{text}. This replaces the entropy-based rule with a criterion that requires both strong alignment to some class anchors and low predictive uncertainty, thereby down-weighting confidently wrong or background-dominated views.

3.2. Image-Guided Semantic Filtering

Our text anchor effectively adapts a semantic prior to the specific view. To complement this semantic-level adaptation, we introduce an image anchor strategy designed to capture the overall visual statistics of the test-time distribution. Specifically, we maintain a class-wise prototype bank that continuously aggregates representative visual features during test-time inference, serving as a memory of how each class appears under the current distribution. From this bank, the image anchor is constructed to provide a compact and dynamically updated representation of the visual domain.

Image Anchor Construction. At test time, views selected by the text-guided filter are used to incrementally update a class-wise prototype bank $\mathcal{P} = \{\mathbf{p}_c\}_{c=1}^{C}$. For each selected embedding $\mathbf{e}$, we obtain its predicted class label c from the original model before adaptation. The corresponding prototype $\mathbf{p}_c$ is updated via cumulative averaging, $\mathbf{p}_c \leftarrow \frac{n_c \mathbf{p}_c + \mathbf{e}}{n_c + 1}$, where n_c is the number of past updates for class c. For a given test image, we compute class scores $\bar{\pi}_c$ by averaging predictions over selected views and choose the top-K class indices, $\mathcal{C}_K = \text{TopK}(\{\bar{\pi}_c\})$. The image anchor set $\mathcal{A}_{\text{img}}$ is then formed by gathering the corresponding prototypes: $\mathcal{A}_{\text{img}} = \{\mathbf{p}_c \mid c \in \mathcal{C}_K\}$.

Joint Alignment-Confidence Scoring. In parallel to the text-guided filtering, we introduce a second stage of view selection driven by image anchors. For all aug views, align-

ment with the image anchor set and predictive confidence are computed similar to Eq. 1. These terms are combined into a single score:

$$\mathrm{S}_{\text{img}}(\mathbf{e}) = \beta_1 s_{\text{align}}(\mathbf{e}) + \beta_2 s_{\text{conf}}(\mathbf{e}), \tag{3}$$

with (β_1, β_2) balancing alignment and confidence. We retain the top-$p\%$ of views with the highest S_{img}. The final set of selected views is the union of both filters, with any duplicates counted once.

3.3. Adaptive Ensemble of Multimodal Predictions

To further enhance stability and accuracy for these selected views, we present a confidence-aware weighting ensemble that dynamically integrates predictions from multiple sources. This approach mitigates the influence of noisy predictions by assigning higher weights to more confident ones, thus generating a refined training signal. The model produces three distinct similarity vectors for each selected view, obtained by computing the cosine similarity between the image feature and (i) the original prompt embeddings, (ii) the text anchors $\mathbf{t}_c$, and (iii) the image prototypes $\mathbf{p}_c$. We denote these as $\mathbf{z}_{\text{prompt}}$, $\mathbf{z}_{\text{text}}$, and $\mathbf{z}_{\text{image}}$, respectively, and refer to the set of sources as $S = \{\text{prompt, text, image}\}$.

To integrate complementary evidence from these sources, we aggregate their outputs based on their predictive confidence. First, we compute the probability distribution for each source as $\mathbf{q}_k = \mathrm{Softmax}(\mathbf{z}_k),\ k \in S$. Next, we define the confidence score γ_k for each source as the maximum probability in its distribution: $\gamma_k = \max(\mathbf{q}_k)$. These confidence scores are then normalized to produce the ensemble weights. The weight w_k for each source $k \in S$ is: $w_k = \frac{\gamma_k}{\sum_{j \in S} \gamma_j + \epsilon}$, where ϵ is a small constant for numerical stability. Finally, the ensemble vector is generated by a weighted sum of the individual scores:

$$\mathbf{z}_{\text{ens}} = \sum_{k \in S} w_k \mathbf{z}_k. \tag{4}$$

This confidence-aware ensemble creates a de-biased supervision signal by forming a robust consensus that leverages the most reliable predictive head for each image, thereby mitigating single-source errors.

Target Distribution Sharpening. We compute per-view probabilities via softmax of the ensemble logits z_{ens}, average it over the selected views, and sharpen it using temperature scaling ($T = 0.3$) to form the target distribution, $\tilde{\mathbf{q}}$. The original prediction, $\mathbf{p}_v$, is formed by averaging the original prompt logits over the same views and applying a softmax function. We then update the prompts by minimizing the KL divergence:

$$\mathcal{L}_{\text{TPT}} = D_{\text{KL}}(\tilde{\mathbf{q}} \,\|\, \mathbf{p}_v). \tag{5}$$

This loss encourages the prompt-driven prediction $\mathbf{p}_v$ to match the confident, ensemble-based target distribution $\tilde{\mathbf{q}}$. Note that prior TPT methods [8, 30, 37] directly minimize the entropy of the averaged distribution, pushing the model toward self-sharpening. Our method stabilizes adaptation by deriving a sharpened target distribution from the ensemble, which prevents collapse into overconfident yet semantically inconsistent predictions.

4. Experiments

4.1. Experimental Setup

Datasets. Following previous methods [8, 30, 37], we evaluate our method in two settings: domain generalization and cross-dataset transfer. For domain generalization, we use ImageNet [5] along with four distribution-shifted test sets: ImageNet-A [11], ImageNet-R [10], ImageNet-V2 [26], and ImageNet-Sketch [33]. For cross-dataset evaluation, we consider 10 diverse image classification benchmarks: Flower102 [21], DTD [4], OxfordPets [22], StanfordCars [14], UCF101 [31], Caltech101 [7], Food101 [3], SUN397 [35], FGVC-Aircraft [18], and EuroSAT [9]. These datasets cover a diverse spectrum of visual domains, ranging from fine-grained object classification to satellite imagery.

Implementation details. Our model is built upon CLIP [25] with a ViT-B/16 backbone. We optimize four learnable prompt tokens in text embedding space. Following TPT [30], we generate 63 random resized crops per test image and include the original image, yielding a batch of 64 views. For each view, we compute entropy and alignment with both text and image anchors, retaining the top 10% for text anchors (6 views) and the top 5% for image anchors (3 views). The final set of selected views is the union of these two sets, resulting in a total of 6 to 9 unique views depending on the degree of overlap. Prompt optimization is performed using KL divergence against the sharpened predictions of the selected augmented views. We use AdamW with a learning rate of 0.003. All experiments were conducted using a single NVIDIA RTX 6000 ada GPU. All prompt tuning methods are trained on ImageNet with 16-shot.

4.2. Main Results

Comparisons on domain generalization. We evaluate our approach under the domain generalization setting by comparing it against both prompt learning and test-time prompt tuning baselines, as summarized in Table 1. The metric *Average* is computed as the mean top-1 accuracy over all five datasets. The metric *OOD Average* is computed as the mean top-1 accuracy over the four out-of-distribution datasets, excluding ImageNet. Prompt learning methods, such as CoOp, CoCoOp, and MaPLe, are trained with the supervised cross-entropy loss on ImageNet, while test-time prompt tuning methods, including TPT, DiffTPT, and Dy-

Table 1. Comparison of methods on multiple ImageNet variants. We report top-1 accuracy (%) on each dataset and their averages. Numbers in parentheses indicate improvement over the corresponding baseline (TPT, CoOp+TPT, or MaPLe+TPT).

Method	ImageNet	ImageNet-A	ImageNet-V2	ImageNet-R	ImageNet-Sketch	Average	OOD Avg.
CLIP-ViT-B/16	66.73	47.87	60.86	73.98	46.09	59.11	57.20
Prompt learning methods without test-time tuning							
CoOp	71.51	49.71	64.20	75.21	47.99	61.72	59.28
CoCoOp	71.02	50.63	64.47	76.18	48.75	62.13	59.91
MaPLe	70.72	49.15	64.07	76.98	50.90	62.36	60.28
CoPrompt	70.80	49.43	64.25	77.51	50.50	62.50	60.42
Any-shift Prompt	–	49.80	64.53	77.56	51.52	–	60.85
Test-time prompt tuning methods							
TPT	68.98	54.77	63.45	77.06	47.06	62.06	60.81
DiffTPT	70.30	55.68	65.10	75.00	46.80	62.26 (+0.20)	60.65 (-0.16)
C-TPT	69.30	52.90	63.40	78.00	48.50	62.42 (+0.36)	60.70 (-0.11)
DynaPrompt	69.61	56.17	64.67	78.17	48.22	63.37 (+1.31)	61.81 (+1.00)
Ours	**72.21**	**59.65**	**65.35**	**80.25**	**51.24**	**65.74 (+3.68)**	**64.12 (+3.31)**
CoOp + TPT	73.61	57.95	66.83	77.27	49.29	64.99	62.84
CoOp + DynaPrompt	**74.08**	60.55	67.25	79.15	50.28	66.26 (+1.27)	64.31 (+1.47)
CoOp + Ours	74.03	**62.40**	**67.50**	**81.02**	**52.58**	**67.51 (+2.52)**	**65.88 (+3.04)**
MaPLe + TPT	71.87	58.08	64.87	78.12	48.16	64.22	62.31
MaPLe + PromptAlign	-	59.37	65.29	79.33	50.23	-	63.56 (+1.25)
MaPLe + DynaPrompt	**72.71**	**60.72**	66.34	79.57	50.25	65.92 (+1.70)	64.22 (+1.91)
MaPLe + Ours	72.31	59.65	**66.84**	**80.50**	**52.11**	**66.28 (+2.06)**	**64.78 (+2.47)**

naPrompt, adapt prompts at inference time using unlabeled test data.

Our method consistently improves performance across all metrics. In particular, we achieve an average accuracy of 65.74 and an OOD average of 64.12, outperforming TPT by +3.68% and +3.04%, respectively. Compared to the strongest baseline, DynaPrompt, our method still provides a clear margin of +2.37% in average and +2.31% in OOD accuracy. Since our method can be seamlessly integrated into existing prompt tuning frameworks, combining it with CoOp and MaPLe further improves performance. Across both bases, our method delivers the best average and OOD average: with CoOp, improvements of +2.52% in average and +3.04% in OOD over TPT; with MaPLe, improvements of +2.06% and +2.47% over TPT. Notably, even without ImageNet-pretrained prompts from CoOp, Ours alone surpasses OOD accuracy of CoOp+TPT, highlighting that our anchor-guided filtering is complementary to test-time prompt tuning.

We further analyze the inference efficiency of our method in comparison with other test-time prompt tuning approaches, as shown in Table 2. The evaluation is performed on the ImageNet-R dataset. While TPT achieves modest improvements with low overhead, its overall gain is limited. DiffTPT attains lower accuracy and already takes over 0.6 sec per image due to multiple prompt-update steps; latency increases further when diffusion is used. DynaPrompt also improves accuracy, yet incurs additional overhead (0.4 sec) due to optimizing and maintaining multiple prompts simultaneously. Our method attains the highest accuracy with negligible overhead (0.2 sec), almost same runtime as vanilla TPT. These results confirm that our anchor-guided filtering introduces minimal computational burden while delivering larger gains.

Table 2. Comparison of test-time tuning methods. Accuracy, gain over the CLIP baseline, and test time are reported.

Method	Accuracy	Gain	Testing Time
CLIP	73.98	-	
TPT	77.06	+3.06	0.2 sec
DiffTPT	75.00	+1.02	$>$ 0.6 sec
DynaPrompt	78.17	+4.19	0.4 sec
Ours	**80.25**	**+6.27**	0.2 sec

Comparisons on cross-dataset. Table 3 presents the top-1 accuracy comparison of our method against prompt learning and test-time prompt tuning methods in 10 downstream datasets. A critical observation from the table is the performance trade-off in test-time tuning. Many existing TPT methods show lower average performance than the best prompt learning methods, such as MaPLe and CoPrompt, that do not perform test-time adaptation. This highlights a significant challenge where test-time adaptation can struggle to improve performance or fail to provide substantial gains over a strong, pretrained prompt. The proposed method successfully overcomes this limitation. It achieves an average accuracy of 68.81%, which not only surpasses all other TPT methods by a large margin but also significantly outperforms the strongest prompt learning method, CoPrompt, by 1.81%. Furthermore, our method demonstrates strong synergy when combined with MaPLe. The MaPLe + Ours variant achieves a new state-of-the-art av-

Table 3. Comparison of methods on multiple downstream datasets.

Method	Flower102	DTD	Pets	Cars	UCF101	Caltech101	Food101	SUN397	Aircraft	EuroSAT	Average
CLIP-ViT-B/16	67.44	44.27	88.25	65.48	65.13	93.35	83.65	62.59	23.67	42.01	63.58
Prompt learning methods without test-time tuning											
CoOp	68.71	41.92	89.14	64.51	65.55	93.70	85.30	58.15	18.47	46.39	63.88
CoCoOp	70.85	45.45	90.46	64.90	68.44	93.79	83.97	66.89	22.29	39.23	64.63
MaPLe	72.23	46.49	90.49	65.57	68.69	93.53	86.20	67.01	24.74	48.06	66.30
CoPrompt	72.30	47.07	90.73	65.67	69.73	94.50	86.43	67.57	24.00	51.90	67.00
Test-time prompt tuning methods											
TPT	68.98	47.75	87.79	66.87	68.04	94.16	84.67	65.50	24.78	42.44	65.10
DiffTPT	70.10	47.00	88.22	67.01	68.22	92.49	**87.23**	65.74	25.60	41.04	65.47 (+0.37)
C-TPT	69.80	46.00	88.20	65.80	65.70	93.60	83.70	64.80	24.00	43.20	64.80 (-0.30)
DynaPrompt	69.95	47.96	88.28	67.65	68.72	94.32	85.42	66.32	24.33	42.28	65.52 (+0.42)
Ours	**74.71**	**53.19**	**89.94**	**68.75**	**72.24**	**94.85**	85.45	**70.28**	**28.11**	**50.62**	**68.81 (+3.71)**
CoOp + TPT	68.48	43.32	89.48	66.77	68.91	93.15	**86.48**	66.06	20.51	37.73	64.09
CoOp + DynaPrompt	69.38	46.98	**90.04**	67.35	69.54	94.40	86.45	66.17	21.35	38.55	65.02 (+0.93)
CoOp + Ours	**74.1**	**49.59**	88.85	**68.83**	**71.24**	**94.60**	84.42	**69.73**	**23.73**	**50.37**	**67.55 (+3.46)**
MaPLe + TPT	72.37	45.87	90.72	66.50	69.19	93.59	86.64	67.54	24.70	47.80	66.50
MaPLe + PromptAlign	72.39	47.24	90.76	68.50	69.47	94.01	86.65	67.54	24.80	47.86	66.92 (+0.42)
MaPLe + DynaPrompt	73.28	48.75	**90.95**	68.26	69.85	**95.17**	86.60	68.18	24.36	47.53	67.29 (+0.79)
MaPLe + Ours	**75.10**	**53.72**	90.32	**69.55**	**72.21**	94.16	**87.41**	**70.32**	**29.11**	**51.69**	**69.36 (+2.86)**

Table 4. Comparison of selection strategies. We compare the conventional confidence-based selection with the proposed anchor-guided selection. The last row reports the performance of our method.

Method	ImageNet	ImageNet-A	ImageNet-V2	ImageNet-R	ImageNet-Sketch	Average	OOD Avg.
CLIP (ViT-B/16)	66.73	47.87	60.86	73.98	46.09	59.11	57.20
TPT (baseline)	68.47	51.12	62.24	75.54	47.20	60.91	59.03
+ Cosine Sel. (Orig. Img.)	69.02	53.67	63.37	77.00	48.00	62.21 (+1.30)	60.51 (+1.48)
+ Confidence Sel.	68.98	54.77	63.45	77.06	47.94	62.44 (+1.53)	60.81 (+1.78)
+ Ours (Anchor Sel.)	**70.53**	**56.83**	**64.54**	**79.07**	**49.05**	**64.00 (+3.09)**	**62.37 (+3.34)**

erage accuracy of 69.38%. This represents a substantial 2.88% improvement over the baseline MaPLe + TPT, proving superior effectiveness of our method as a test-time tuning strategy.

4.3. More Studies

Comparisons of selection strategies. We compare several view selection approaches to examine their impact on test-time prompt tuning. *TPT (baseline)* does not apply any selection mechanism. *Cosine Sel. (Orig. Img.)* incorporates both the prediction entropy and the cosine similarity to the original image as selection criteria. *Confidence sel* only incorporates prediction entropy, selecting views that show low prediction entropy. The cosine selection method showed the lowest performance, as shown in Table 4, demonstrating that simple similarity to the original image alone cannot serve as a reliable semantic indicator. Our *Anchor Sel.* attains 64.00% in average and 62.37% in OOD average accuracy, outperforming TPT by +3.09% and +3.34%, respectively. This demonstrates that effective view selection is key to TPT.

Effectiveness of Text and Image Anchors. To analyze the contribution of each component in our dual-modality framework, we conduct an ablation study by incrementally activating the text and image anchors for both filtering and ensemble stages, as summarized in Table 5. Starting from the baseline TPT [30] (✗ in all rows), introducing text anchor filtering (S_{text}) improves accuracy across all datasets, verifying that text anchor helps retain informative views. Adding image anchor filtering (S_{image}) leads to additional improvements, confirming that image anchor provides complementary cues for view quality. When we additionally incorporate text anchor predictions (z_{text}) and image anchor predictions (z_{image}) in the ensemble, performance consistently improves. The full configuration achieves the best results, highlighting the effectiveness of the proposed dual-modality anchor-guided framework.

Table 5. Ablation study on four components and the results on three benchmarks.

$\mathbf{S}_{\text{text}}$	$\mathbf{S}_{\text{image}}$	$\mathbf{z}_{\text{text}}$	$\mathbf{z}_{\text{image}}$	Flowers	DTD	ImageNet
✗	✗	✗	✗	69.39	46.63	68.98
✓	✗	✗	✗	70.36	49.41	70.52
✗	✓	✗	✗	70.08	47.75	70.01
✓	✓	✗	✗	71.54	49.70	70.53
✓	✓	✓	✗	72.84	51.36	71.04
✓	✓	✗	✓	72.76	51.47	71.34
✓	✓	✓	✓	**74.71**	**53.19**	**72.21**

Sensitivity to hyperparameters. We study the influence of the weighting ratios in Eqs. 2 and 3 that control the

Table 6. Effect of hyperparmeters.

$\alpha_1 : \alpha_2$	1:4	1:2	1:1	2:1	4:1
OOD Avg (%)	62.73	**63.01**	62.95	62.87	62.66

$\beta_1 : \beta_2$	1:4	1:2	1:1	2:1	4:1
OOD Avg (%)	62.75	62.83	62.95	**63.01**	62.80

Table 7. Effect of loss and ensemble strategy on cross-dataset (CD) and out-of-distribution (OOD) performance.

Setting	OOD (%)	CD (%)
EM + Simple Avg.	61.15	66.72
EM + Conf. Ens.	62.23	67.56
KLD + Simple Avg.	62.51	67.53
KLD + Conf. Ens. (Ours)	**63.01**	**68.81**

trade-off between the alignment and entropy terms in our joint scoring modules. Results in Table 6 show that the text-anchor module attains its best accuracy at 1:2 (alignment:entropy). The image-anchor module is optimal at 2:1.

Effectiveness of alignment-weighted aggregation. We investigate the impact of alignment-weighted aggregation, which adaptively fuses multiple LLM-generated text descriptions into a single anchor based on their semantic relevance to the image. Without aggregation, the anchor is formed by simply averaging all description embeddings. The simple averaged text anchor achieves 62.53% OOD accuracy, while the alignment-weighted aggregated text anchor achieves 63.01%. Incorporating alignment-weighted aggregation thus improves OOD accuracy by 0.48%, confirming that weighting descriptions by their alignment scores yields more faithful and informative anchors.

Effect of loss design and ensemble strategy. We analyze the contribution of our sharpened KLD loss and the confidence-aware ensemble used during test-time adaptation. As shown in Table 7, replacing the standard entropy minimization (EM) loss with the KLD objective improves both cross-dataset (CD) and out-of-distribution (OOD) accuracy. While simple averaging of predictions provides marginal gains, incorporating confidence-aware weighting yields further improvements by stabilizing the target distribution. The combination of the KLD loss and confidence-aware ensemble achieves the best overall performance, showing that the two components are complementary.

4.4. Qualitative Analysis of View Selection

To better understand the role of each selection component, we qualitatively compare the augmented views retained by different filtering strategies in Fig. 3. Entropy refers to the entropy-based selection used in prior TPT methods [30, 37, 39]. Cosine refers to the selection strategy [8] which incorporates both prediction entropy and cosine similarity to the original image. Entropy-based selection often focuses on background or misaligned regions, revealing its tendency to prioritize low-entropy predictions rather than semantically meaningful ones. Similarly, cosine-based selection often favors views that are visually similar to the original image but semantically misaligned, revealing its limitation in capturing meaningful object-level consistency.

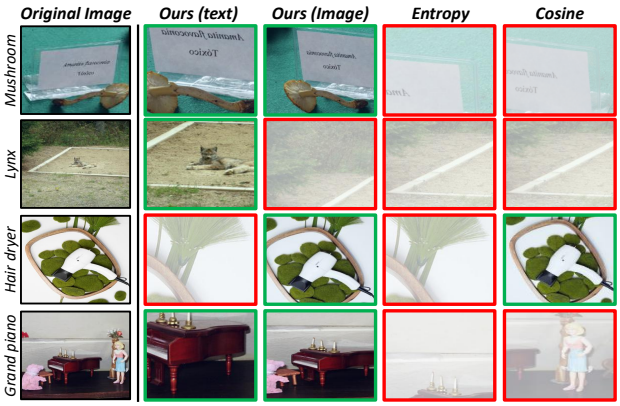

Figure 3. Visualization of views selected by each anchor modality. For each test image, we compare the augmented views selected by the text-guided and image-guided filters with the previous selection strategies. Green boxes indicate correctly selected views containing the target object, whereas red boxes mark views where the object is absent or unrecognizable.

Text anchor filtering in the proposed method alleviates this issue by emphasizing semantically faithful and class-discriminative regions aligned with textual descriptions. Image anchor filtering complements the text-guided counterpart by selecting visually coherent and content-preserving views, guided by image anchors that capture the underlying distribution of test samples. With the text and image filters , the method recovers cases where one modality alone fails.

5. Conclusion

We present a dual-modality anchor framework for test-time prompt tuning that grounds adaptation in both semantic and visual evidence. Unlike prior entropy–based filtering approaches, our method constructs semantically enriched text anchors through alignment-weighted aggregation and complements them with incrementally updated image anchors derived from a prototype bank. The dual anchors jointly provide alignment–confidence scoring that balances semantic fidelity and uncertainty, yielding more reliable view selection. Moreover, these anchors also serve as predictive sources whose ensemble provides a reliable target distribution, leading to stable adaptation. Extensive experiments across multiple datasets demonstrate that our framework consistently outperforms existing test-time tuning methods, achieving robust performance.

Acknowledgement. This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)]

References

[1] Jameel Abdul Samadh, Mohammad Hanan Gani, Noor Hussein, Muhammad Uzair Khattak, Muhammad Muzammal Naseer, Fahad Shahbaz Khan, and Salman H Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. *Advances in Neural Information Processing Systems*, 36:80396–80413, 2023. 2, 3

[2] James Urquhart Allingham, Jie Ren, Michael W Dusenberry, Xiuye Gu, Yin Cui, Dustin Tran, Jeremiah Zhe Liu, and Balaji Lakshminarayanan. A simple zero-shot prompt weighting technique to improve prompt ensembling in text-image models. In *International Conference on Machine Learning*, pages 547–568. PMLR, 2023. 1

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014. 5

[4] 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. 5

[5] 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. Ieee, 2009. 5

[6] Yingjun Du, Wenfang Sun, and Cees Snoek. Ipo: Interpretable prompt optimization for vision-language models. *Advances in Neural Information Processing Systems*, 37: 126725–126766, 2024. 1

[7] 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. 5

[8] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2704–2714, 2023. 2, 3, 5, 8

[9] 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. 5

[10] 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*, pages 8340–8349, 2021. 5

[11] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271, 2021. 5

[12] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1, 2, 3

[13] 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, 3

[14] 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. 5

[15] 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. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[16] Xiaokun Li, Yaping Huang, and Qingji Guan. Clip-driven coarse-to-fine semantic guidance for fine-grained open-set semi-supervised learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30312–30321, 2025. 2

[17] Chengcheng Ma, Yang Liu, Jiankang Deng, Lingxi Xie, Weiming Dong, and Changsheng Xu. Understanding and mitigating overfitting in prompt tuning for vision-language models. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4616–4629, 2023. 1

[18] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 5

[19] Mayug Maniparambil, Chris Vorster, Derek Molloy, Noel Murphy, Kevin McGuinness, and Noel E O'Connor. Enhancing clip with gpt-4: Harnessing visual descriptions as prompts. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 262–271, 2023. 3

[20] Muhammad Ferjad Naeem, Muhammad Gul Zain Ali Khan, Yongqin Xian, Muhammad Zeshan Afzal, Didier Stricker, Luc Van Gool, and Federico Tombari. I2mvformer: Large language model generated multi-view document supervision for zero-shot image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15169–15179, 2023. 3

[21] 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. 5

[22] 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. 5

[23] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15691–15701, 2023. 3

[24] Xiangyan Qu, Gaopeng Gou, Jiamin Zhuang, Jing Yu, Kun Song, Qihao Wang, Yili Li, and Gang Xiong. Proapo: Progressively automatic prompt optimization for visual classification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25145–25155, 2025. 1, 3

[25] 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, 2, 3, 5

[26] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 5

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[28] Shuvendu Roy and Ali Etemad. Consistency-guided prompt learning for vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2

[29] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, 2020. 2

[30] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022. 1, 2, 3, 5, 7, 8

[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. 5

[32] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 1

[33] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in neural information processing systems*, 32, 2019. 5

[34] Shuoyuan Wang, Yixuan Li, and Hongxin Wei. Understanding and mitigating miscalibration in prompt tuning for vision-language models. *arXiv preprint arXiv:2410.02681*, 2024. 2

[35] 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. 5

[36] Zehao Xiao, Jiayi Shen, Mohammad Mahdi Derakhshani, Shengcai Liao, and Cees GM Snoek. Any-shift prompting for generalization over distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13849–13860, 2024. 1, 2

[37] Zehao Xiao, Shilin Yan, Jack Hong, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiayi Shen, Qi Wang, and Cees GM Snoek. Dynaprompt: Dynamic test-time prompt tuning. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 3, 5, 8

[38] Zhiwei Yang, Yucong Meng, Kexue Fu, Feilong Tang, Shuo Wang, and Zhijian Song. Exploring clip's dense knowledge for weakly supervised semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20223–20232, 2025. 2

[39] Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang D Yoo. C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In *12th International Conference on Learning Representations, ICLR 2024*, 2024. 2, 3, 8

[40] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 1

[41] 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. 1, 2

[42] 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

[43] Yuhan Zhu, Yuyang Ji, Zhiyu Zhao, Gangshan Wu, and Limin Wang. Awt: Transferring vision-language models via augmentation, weighting, and transportation. *Advances in Neural Information Processing Systems*, 37:25561–25591, 2024. 3