IMA & TMA: Efficient Test-Time Adaptation for VLMs via Linear Transformation in Embedding Space

Rishik Vamshi Rohith Vempati* Eswar Venkata Sai Kadava* Konda Reddy Mopuri
Department of Artificial Intelligence, Indian Institute of Technology Hyderabad
{ai24mtech12003@iith.ac.in, ai24mtech11007@iith.ac.in, krmopuri@ai.iith.ac.in}

Abstract

Large-scale Vision-Language Models (VLMs) have set new benchmarks in zero-shot learning; however, their performance remains brittle under distribution shifts at test time. While existing Test-Time Adaptation (TTA) methods often rely on prompt tuning or input-space optimization, they incur significant computational overhead and scale poorly with class cardinality. To bridge this gap, we propose two lightweight, sample-wise alignment strategies: Image Matrix Adapter (IMA) and Text Matrix Adapter (TMA). Unlike previous methods, IMA and TMA apply linear corrections directly in the embedding space, thereby restoring cross-modal alignment with a single test sample. This approach drastically reduces memory and computational requirements, as the adaptation cost remains independent of the number of target classes. Extensive evaluations across diverse out-of-distribution (OOD) benchmarks and cross-dataset scenarios demonstrate that our methods achieve competitive accuracy while being significantly more efficient than state-of-the-art prompt-based adaptation, making them ideal for resource-constrained deployment. The code can be found here.

1. Introduction

Vision–language foundation models (VLMs) [27, 34, 51] have significantly advanced computer vision. Pretrained on web-scale image–caption pairs with contrastive learning, they encode diverse visual concepts through language supervision. Consequently, VLMs have become the de facto choice for numerous downstream tasks [46, 53], in a zero-shot setting with out task specific training.

Despite being trained on massive databases [36, 37], VLMs still suffer significant performance drops under train–test distribution shifts. They rely on domain-specific prompts to properly align modalities, yet designing such prompts is challenging as handcrafted templates are subjective and often suboptimal, even when assisted by large language models. To mitigate these issues, recent work adapts VLMs through prompt learning and fine-tuning. In particular, prompt tuning [24, 57, 58] learns continuous prompt embeddings from task-specific training data by optimizing differentiable prompt representations. While such learned prompts often outperform handcrafted ones, they remain closely tied to the training distribution and task characteristics, limiting generalization beyond the source domain. Moreover, prompt tuning and other fine-tuning approaches require annotated data, making them unsuitable for zero-shot or deployment-time settings where labels are unavailable or costly. Beyond prompt learning, adaptation strategies ranging from full-model fine-tuning to parameter-efficient methods [19, 22, 26] have been explored for improving VLM performance. However, these approaches require labeled data and may suffer from catastrophic forgetting, limiting their applicability in annotation-scarce test domains. To address these challenges, we focus on the emerging paradigm of test-time adaptation (TTA) [4, 5, 28, 48], which dynamically adapts pretrained models at inference using only unlabeled test samples from an unknown target distribution in a self-supervised manner.

Among various TTA settings, including TTBA (Test-Time Batch Adaptation), OTTA (Online Test-Time Adaptation), and TTDA (Test-Time Domain Adaptation), episodic TTA has recently gained attention as a special case of TTBA with batch size one. Unlike OTTA, which adapts sequentially and may accumulate noise across unrelated samples, it resets the model to its pretrained state for each test instance. In contrast to TTDA, which assumes access to the entire target dataset and operates offline, it is better suited for realistic deployment where test samples arrive independently.

In this work, we propose two simple episodic matrix-based adaptation techniques, IMA and TMA. IMA learns a linear transformation on the image embedding while keeping textual prototypes fixed, whereas TMA learns a transformation on the textual prototypes while keeping the image embedding fixed. Both methods perform single-sample adaptation by modifying one modality to better bridge the domain gap. Importantly, backpropagation is restricted to the embed-

*Equal contribution

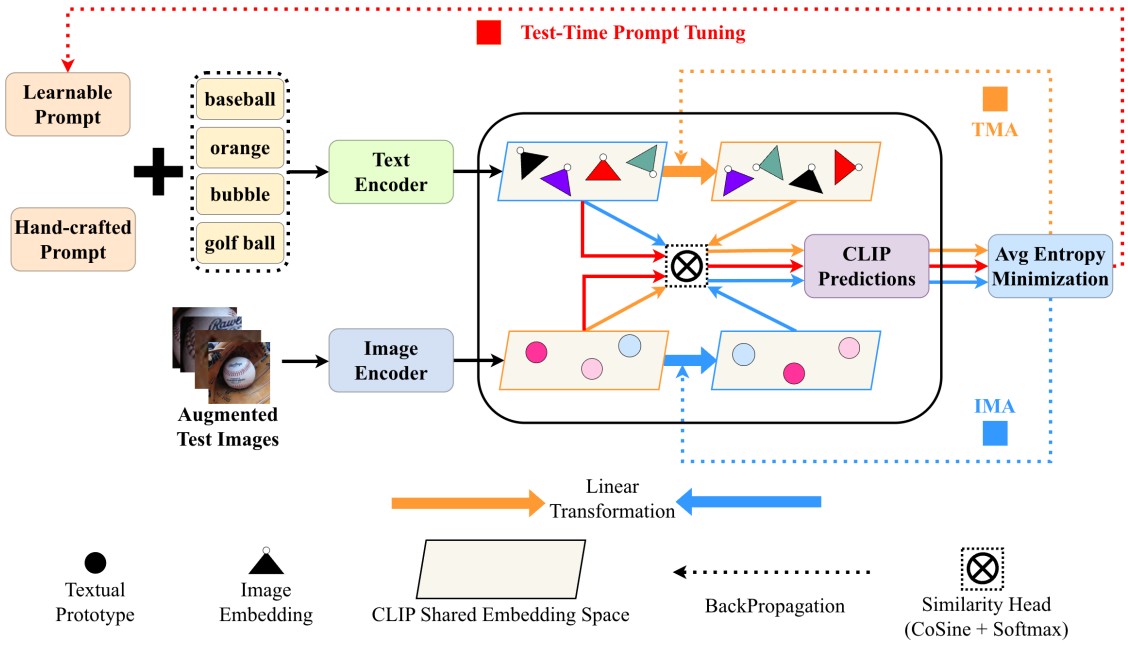


Figure 1. Overview of the proposed framework for test-time adaptation process within the shared embedding space. While TPT [39] backpropagates gradients through the large text encoder to optimize learnable prompt, whereas our method uses vanilla Hand-crafted textual prompt and performs adaptation directly in the embedding space by learning a lightweight transformation matrix. This restricts gradient computation to the embedding space, resulting in significantly lower memory and computational overhead.

ding space, enabling efficient updates of the transformation parameters without propagating gradients through the heavy encoders.

Our main contributions are summarized as follows:

1. We introduce **linear transformation-based adaptation frameworks** that operate directly in the shared embedding space, enabling cross-modal alignment at test time by adapting either of the modalities.
2. The proposed transformations are **lightweight**, with complexity depending only on the embedding dimensionality rather than the number of target classes, unlike existing TTA methods such as TPS [41].
3. Our approach achieves significant reduction in computation and memory vs. TPT and TTL [21], while being on par, even slightly memory efficient to peer-embedding based approach TPS (with ViT-B/32 [8] backbone).

2. Related Work

2.1. Test-Time Adaptation

Test-time adaptation (TTA) aims to adapt a pretrained model to samples from an unknown target distribution during inference without access to labels or source data. Early work such as TENT [44] proposed entropy minimization to adapt Batch Normalization parameters under distribution shifts. MEMO [54] extends this idea to episodic settings by adapting network parameters through marginal entropy minimization across augmented views of a single test sample.

For VLMs, TPT [39] performs episodic adaptation by optimizing textual prompt tokens using entropy minimization. Several variants further improve this idea: DiffTPT [12] introduces diffusion-based augmentations for more semantically consistent views, C-TPT [49] enhances prediction calibration by maximizing average text feature dispersion (ATFD), and R-TPT [38] improves adversarial robustness by refining the entropy objective and weighting augmented views based on reliability. PromptAlign [1] addresses domain shift through multimodal prompt alignment with proxy source statistics, while TTL [21], inspired by LoRA [20], inserts learnable low-rank matrices into the visual encoder. Moving adaptation to the feature space, TPS [41] learns shift vectors for textual prototypes directly in the embedding space, reducing computational cost but requiring parameters that scale with the number of target classes. TACT [29] trims out non-causal components using PCA, assuming classification-relevant information lies in causal directions.

Another line of work studies online TTA [2, 13, 23, 52, 55, 59] by leveraging historical test samples. While these approaches follow different strategies like cache-based logits, residual prototype updates, or statistical memory to improve adaptation by exploiting accumulated information from prior test instances. However, these methods depend on the order of the test stream, where early blocking samples can corrupt the cache and propagate noisy predictions.

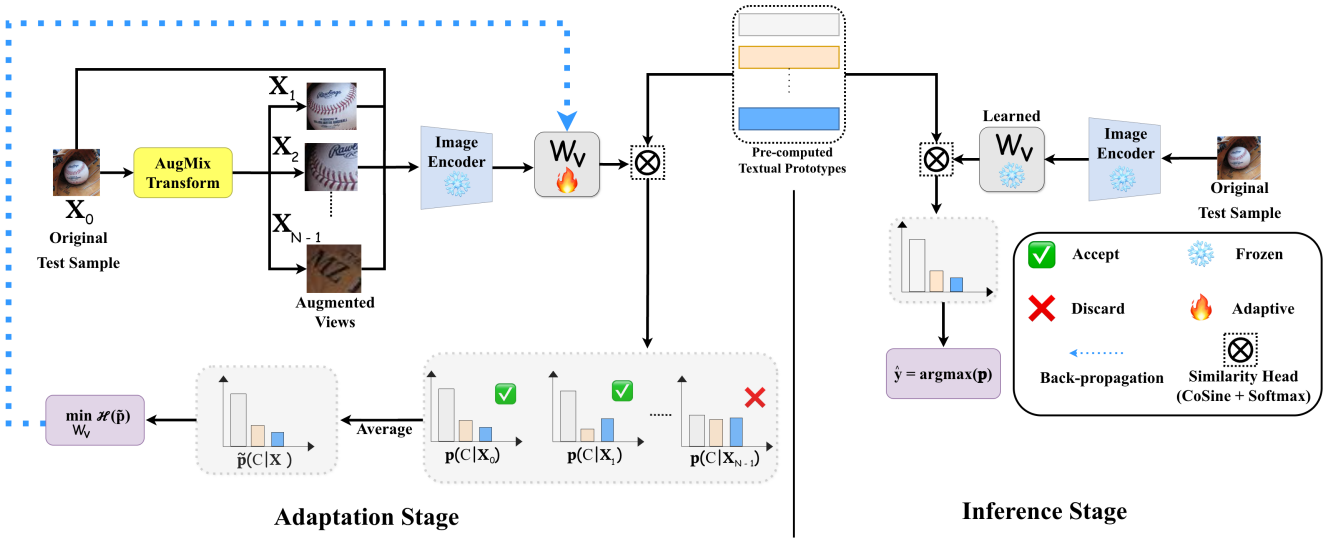


Figure 2. Workflow of **Image Matrix Adapter(IMA)**. **1) Test-Time Matrix Adaptation**: For a given test image, multiple augmented views are generated and their CLIP image embeddings are computed. Augmentations with low zero-shot entropy are selected as confident views. A lightweight transformation matrix W_v is then optimized in the embedding space by minimizing the marginal entropy of CLIP similarities between the fixed textual class prototypes and initial transformation applied on the image embeddings. This adapts the image representation to reduce the domain gap without updating the encoders. **2) Test-Time Inference**: Using the learned transformation matrix W_v, the embedding of the original test image is transformed and compared with the fixed class prototypes using CLIP similarity. The final prediction is obtained from **argmax** of the similarity scores of the transformed original image embedding and textual prototypes.

2.2. Prompt Learning for TTA

Foundational models such as CLIP [34] encode rich knowledge that can be transferred to downstream tasks in a zero-shot setting through prompting. However, performance is highly sensitive to prompt design, as manually crafted prompts are often suboptimal. Prompt tuning addresses this by learning task-specific prompts as a parameter-efficient alternative to full fine-tuning. For instance, prompt ensembling [34] aggregates predictions from multiple prompts for improved robustness, while CoOp [58] learns continuous prompt representations from few-shot data to enhance performance. CoCoOp [57] further conditions prompts on input instances to improve robustness under distribution shifts.

3. Methodology

3.1. CLIP Background

CLIP (Contrastive Language–Image Pretraining) [34] is a two-tower architecture consisting of image and text encoders trained with a contrastive objective on roughly 400M image–caption pairs. The objective aligns embeddings of positive image–text pairs while pushing apart negative pairs in a shared embedding space. As a result, classification can be performed through vision–language similarity rather than a fixed classifier head, enabling open-vocabulary and zero-shot recognition without task-specific fine-tuning.

3.2. CLIP for Zero-Shot Image Classification

Let V denote the input image and $\{c_i\}_{i=1}^{K}$ the set of K target classes. In CLIP, the visual encoder f_v maps the image to a d-dimensional embedding $f_v(V)$, while the text encoder f_t maps natural language prompts to the same embedding space. A predefined prompt p (e.g., "a photo of a") is prepended to each class name to form descriptions $\{p_{c_i}\}_{i=1}^{K}$, which are encoded as textual prototypes $\mathbf{t}_{c_i} = f_t(p_{c_i})$. Classification is performed by computing cosine similarities between $f_v(V)$ and $\mathbf{t}_{c_i}$, followed by a softmax with temperature τ; the class with the highest probability is inferred as the predicted label.

$$P(y = c_i \mid V) = \frac{\exp(\text{sim}(f_v(V), \mathbf{t}_{c_i})/\tau)}{\sum_{k=1}^{K} \exp(\text{sim}(f_v(V), \mathbf{t}_{c_k})/\tau)} \quad (1)$$

$$\hat{y} = \arg\max_{c_i} P(y = c_i \mid V) \quad (2)$$

3.3. Linear Transformation for Feature Adaptation

Both proposed variants, IMA and TMA, consist of the following three main stages.

3.3.1. Textual Prototype Generation

In this stage, we use the vanilla prompt template p_{class}, *"a photo of a [class]"*, to generate class descriptions and precompute the textual embeddings $\{\mathbf{t}_{c_i} \in \mathbb{R}^d\}_{i=1}^{K}$ using the CLIP text encoder, where d denotes the shared embedding

dimensionality. These embeddings are cached prior to adaptation. Unlike prompt-learning methods such as TPT [39], our embedding-space adaptation leaves text encoder outputs unchanged, allowing cached prototypes to be reused.

3.3.2. Test-Time Matrix Adaptation

During this stage, we learn a lightweight linear correction matrix to mitigate domain shift caused by misalignment in the shared embedding space. In the IMA variant, a matrix $W_v \in \mathbb{R}^{d\times d}$ is applied to image embeddings, while in TMA a matrix $W_t \in \mathbb{R}^{d\times d}$ transforms textual prototypes. Both matrices are initialized as identity. Cosine similarities are computed between transformed embeddings and textual prototypes across confident augmented views, producing prediction distributions for each view. The matrix (W_v or W_t) is then optimized by minimizing the entropy of the combined marginal distribution, encouraging consistent predictions across augmentations without relying on pseudo-labels.

3.3.3. Test-Time Inference

Here, the learned transformation matrix is applied to the corresponding modality. In IMA, W_v transforms the image embedding, while in TMA, W_t transforms the textual prototypes. Cosine similarities between the transformed modality (IMA: image embedding, TMA: textual prototypes) and its unchanged counterpart produce class-wise scores, followed by softmax to get the final probability distribution. The class with the highest probability is selected as the predicted label.

3.4. IMA for Image Classification

The objective of this variant is illustrated in Fig. 2 is to learn a simple linear correction matrix W_v that is applied to the image modality in the shared embedding space to mitigate the domain shift. Given an unlabeled test sample x_0 with its corresponding image embedding v_0, we first generate $(N-1)$ randomly augmented versions $\{x_j\}_{j=1}^{N-1}$ using the augmentation strategy described in AugMix [16], resulting in a total of N image instances $\{x_j\}_{j=0}^{N-1}$, including the original sample. Following a strategy similar to TPT [39], we aim to filter out noisy augmentations that may not preserve class-discriminative information by introducing a *confidence-based selection* mechanism. Specifically, we retain only those augmented samples whose zero-shot CLIP predictions exhibit low self-entropy (i.e., high confidence). Using the masking criterion $\mathbb{1}\{H(x_j) \leq \tau\}$, we select samples whose prediction entropy H falls below a threshold τ. Here, τ is an instance-specific adaptive threshold defined as the entropy value at $\rho - percentile$ of the self-entropy distribution computed over N augmented views, ranked from low to high (i.e., from high to low confidence).

Let $\{x'_j\}_{j=0}^{s-1}$ denote the subset of the augmented views that satisfy the masking criterion, and let $\{v'_j\}_{j=0}^{s-1}$ represent their corresponding image embeddings obtained by passing through the image encoder. We then pre-multiply the learnable matrix W_v with these filtered embeddings to obtain the corrected image representations of image embeddings $\{W_v v'_j\}_{j=0}^{s-1}$. These transformed embeddings are aligned with fixed textual prototypes by computing cosine similarity between each $W_v v'_j$ and K pre-computed textual embeddings, followed by a softmax operation with temperature parameter τ is applied to get s probability distributions. During the adaptation phase, the matrix W_v is optimized by minimizing the Shannon entropy of the resulting marginalized distribution, thereby encouraging confident and consistent predictions across the selected augmented views.

$$L = -\sum_{i=1}^{K} \tilde{p}(c_i \mid (\mathbf{x}_0, W_v), t_{c_i}) \cdot \log\left(\tilde{p}(c_i \mid (\mathbf{x}_0, W_v), t_{c_i})\right) \tag{3}$$

$$where\ \tilde{p}(c_i \mid (x_0, W_v), t_{c_i}) = \frac{1}{s}\sum_{j=0}^{s-1} p(c_i \mid (x'_j, W_v), t_{c_i}) \tag{4}$$

$$\text{and } p(c_i \mid (x'_j, W_v), t_{c_i}) = \frac{\exp(\text{sim}((W_v v'_j), t_{c_i})/\tau)}{\sum_{k=1}^{K}\exp(\text{sim}((W_v v'_j), t_{c_k})/\tau)} \tag{5}$$

The objective encourages consistent and confident predictions across multiple augmented views while restricting the adaptation to the entries of the linear correction matrix W_v which is the only learnable component throughout the process. Once W_v is optimized, it is applied during the inference phase by pre-multiplying with the original image embedding v_0 to obtain the corrected representation. Cosine similarity is then computed between this transformed embedding and all fixed class-specific textual prototypes to produce the final probability distribution over the target classes. The class corresponding to the highest probability i.e., the top-1 prediction obtained via the **argmax** of the logits, is selected as the final inferred label.

$$\hat{y} = \arg\max_{c_i} p(y = c_i \mid (x_0, W_v^*), t_{c_i}) \tag{6}$$

where W_v^* represents adapted image matrix

3.5. TMA for Image Classification

The objective of this variant, as illustrated in Fig. 3 is to learn a simple linear transformation W_t that is applied to the text modality in the shared embedding space to mitigate the domain gap. Given an unlabeled test sample x_0 with its corresponding image embedding v_0, let $\{x_j\}_{j=0}^{N-1}$ denote the set of all augmented views including the original sample, and $\{x'_j\}_{j=0}^{s-1}$ be the subset selected using the same confidence based filtration procedure as in IMA variant, with $\{v'_j\}_{j=0}^{s-1}$ representing their corresponding embeddings. We then pre-multiply the initially unlearned matrix W_t with each of the pre-computed textual prototypes to obtain corrected prototypes $\{W_t t_{c_i}\}_{i=1}^{K}$. These fixed, filtered image embeddings are aligned with the transformed textual prototypes,

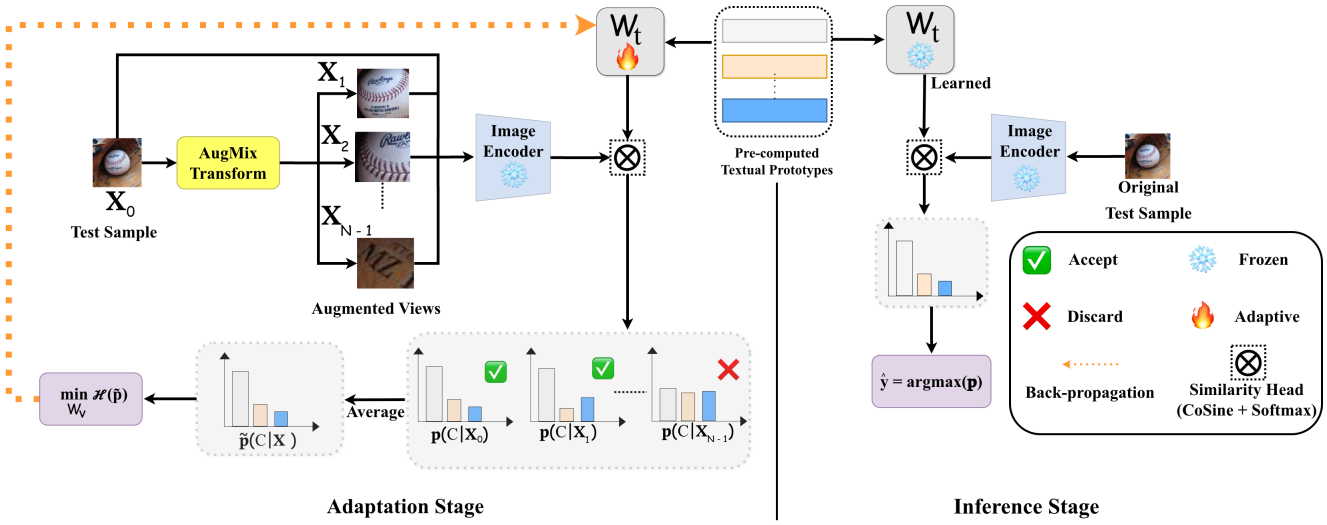


Figure 3. Workflow of **Test Matrix (TMA)**. **1) Test-Time Matrix Adaptation**: Given a test image, multiple augmented views are generated and their CLIP image embeddings are computed. Augmentations with low zero-shot entropy are selected as confident views. A transformation matrix W_t is optimized to adapt the textual class prototypes in the embedding space by minimizing the marginal entropy of CLIP similarities between image embeddings and the initial transformation applied on the textual prototypes. **2) Test-Time Inference**: Using the learned transformation matrix W_t, the textual prototypes are transformed while keeping the original image embedding fixed. The final prediction is obtained from **argmax** of similarity scores between the image embedding and the transformed textual prototypes.

followed by a softmax operation to produce s probability distributions. As in IMA, the matrix W_t is optimized during the adaptation phase by minimizing the Shannon entropy of the resulting marginalized distribution, thereby encouraging invariance in CLIP predictions across the selected augmented views. During the inference phase, the learned matrix W_t is pre-multiplied with all pre-computed textual prototypes, and cosine similarity between the original image embedding v_0 and these transformed prototypes is used to obtain the final class probability distribution, from which prediction is determined by the **argmax** operation.

$$\hat{y} = \arg\max_{c_i} p(y = c_i \mid x_0, (t_{c_i}, W_t^*)) \quad and \tag{7}$$

$$p(c_i \mid x_0, (t_{c_i}, W_t^*)) = \frac{\exp(\text{sim}(v_0, (W_t^* t_{c_i}))/\tau)}{\sum_{k=1}^{K} \exp(\text{sim}(v_0, (W_t^* t_{c_k})/\tau)} \tag{8}$$

where W_t^* represents adapted test matrix.
Algorithms of IMA, TMA are in Section 7 of supplementary.

3.6. Why Linear Transformation?

Let $v_0 \in R^d$ denote the image embedding of a test sample and $\{t_{c_i}\}_{i=1}^K$ the textual prototypes for the K target classes. In IMA, we learn an image-side linear correction $W_v \in R^{d\times d}$, producing the normalized adapted embedding

$$\overline{v_0} = \frac{W_v v_0}{||W_v v_0||_2}, \tag{9}$$

whereas in TMA we learn a text-side linear correction $W_t \in R^{d\times d}$, yielding normalized adapted textual prototypes

$$\overline{t_{c_i}} = \frac{W_t t_{c_i}}{||W_t t_{c_i}||_2}, i = 1, ..., K. \tag{10}$$

Under the normalized CLIP embedding space, which can be viewed as d-dimensional unit sphere, both variants adapt a single modality so that it moves on the sphere to better align with the other, fixed modality. We hypothesize that minimizing the marginal entropy encourages this corrected modality to move toward the ground-truth direction of its complementary modality, thereby improving cross-modal alignment under domain shift. Linear Feature alignment is also supported in the domain adaptation literature. For instance, classical methods such as CORAL [42] and its deep variants demonstrate that linear transformation can effectively reduce domain discrepancy by aligning feature statistics. More recent CLIP-based approaches, including LADS [9] and TADA [15], further indicate that domain shifts in vision-language models often manifest as structured geometric changes in the embedding space, which can be mitigated through linear corrections, at least under simplified shift scenarios. Additionally, the number of learnable parameters in our formulation is fixed at d^2 making the memory overhead independent of the number of target classes.

4. Experimental Results

This section presents the comprehensive evaluation protocol, tasks, and benchmarks used to assess the effectiveness of our proposed linear feature transformation variants, IMA and TMA, which are specifically designed for single-sample

episodic test-time adaptation. We primarily focus on evaluating the model's generalization capability and report results on classification benchmarks that measure robustness to natural distribution shifts as well as cross-dataset generalization under our adaptation framework.

4.1. Datasets

Following the evaluation protocol of [34], we assess robustness to natural distribution shifts using four ImageNet-based OOD benchmarks relative to ImageNet [7]. **ImageNet-V2** [35] provides a re-collected test set of 10k images reflecting distribution shifts. **ImageNet-A** [18] contains 7.5k naturally occurring adversarial examples across 200 classes. **ImageNet-R** [17] includes 30k artistic renditions of ImageNet categories, and **ImageNet-Sketch** [45] consists of 50k sketches evaluating shape-based generalization.

To evaluate cross-dataset generalization, we further test on ten diverse benchmarks spanning multiple visual domains: fine-grained recognition (**Flowers102** [32], **OxfordPets** [33]), transportation (**StanfordCars** [25], **FGVC-Aircraft** [31]), scenes (**SUN397** [47]), textures (**DTD** [6]), food (**Food101** [3]), actions (**UCF101** [40]), satellite imagery (**EuroSAT** [14]), and general object recognition (**Caltech101** [11]).

4.2. Implementation Details

We adopt CLIP [34] with a ViT-B/32 [8] image encoder and Transformer [43] text encoder. For each test sample, we generate 63 augmented views using AugMix [16], resulting in 64 views including the original image. A confidence selection module is applied to retain views corresponding to bottom 10^{th} percentile ($\rho = 0.1$) for MEM in terms of their entropy. The softmax temperature τ is set to 0.07. We optimize using AdamW [30] with learning rates 5×10^{-3} for ImageNet OOD benchmarks and 1×10^{-3} for cross-dataset evaluation. The transformation matrices W_v (IMA) and W_t (TMA) are initialized as identity, preventing the adapted embeddings to largely deviate from pretrained representation during optimization, both of which tuned for a single adaptation step. Class prototypes are built using the vanilla prompt template "*a photo of a*" followed by the class name, and Top-1 accuracy is reported.

4.3. Baselines

We compare our approach with standard zero-shot inference using vanilla prompt template as well as several existing Back-propgataion based TTA methods, all built upon CLIP ViT-B/32 [8] backbone, which serves as our primary baseline. The evaluated TTA approaches include episodic methods such as TPT [39], C-TPT [49], R-TPT [38], which performs textual prompt tuning at test time; TPS [41], which shifts textual prototypes in embedding space; TTL [21] which adapts in encoder space by introducing low rank matrices.

4.4. Aggregation strategy for Augmented Views

For both IMA and TMA, we employ a confidence selection module by default, where only augmented views with low prediction entropy are used to construct the marginal distribution for entropy minimization. This filtered aggregation suppresses unreliable views under distribution shift and stabilizes adaptation. We also evaluate a TTL [21] inspired variant that retains all augmented views. Instead of marginal aggregation, it minimizes a confidence-weighted entropy objective over individual views, where weights derived from zero-shot entropy assign higher importance to low-entropy views and vice versa. The weights remain fixed while optimizing the transformation matrix (W_v for IMA and W_t for TMA), with all other components unchanged. We denote default variant as IMA/TMA + Filtered Augmentations and weighted variant as IMA/TMA + All Augmentations.

4.4.1. Adaptation Phase For Weighted Variant

$$L(W) = \frac{1}{N} \sum_{j=0}^{N-1} \alpha_j H_j(W) \tag{11}$$

$$H_j(W) = -\sum_{i=1}^{K} p(c_i \mid (x_j, t_{c_i}, W)) \log(p(c_i \mid (x_j, t_{c_i}, W)) \tag{12}$$

where H_j is per view entropy loss $p(c_i \mid (x_j, t_{c_i}, W))$ can be computed according to (5) for IMA and (8) for TMA variant respectively.

Here, the weights α_j for each of augmented view x_j are computed from its zero-shot prediction entropy as

$$\alpha_j = \frac{1}{\exp(H_j^{\text{zs}} - \epsilon)}, \quad \epsilon = 0.4. \tag{13}$$

$$H_j^{zs} = -\sum_{i=1}^{K} p(c_i \mid (x_j, t_{c_i})) \log(p(c_i \mid (x_j, t_{c_i}))) \tag{14}$$

where H_i^{zs} is per-view zero-shot prediction entropy.

4.4.2. Inference Phase For Weighted Variant

The label can be inferred using original test sample x_0 and adapted transformation W^* similar to default variant as

$$\hat{y} = \arg\max_{c_i} p(c_i | (x_0, t_{c_i}, W^*)) \tag{15}$$

4.5. Results

4.5.1. Natural Distribution Shifts

Table 1 reports Top-1 accuracy on ImageNet and its OOD variants compared with zero-shot CLIP [34] and prior TTA methods using the ViT-B/32 [8] backbone. Our lightweight embedding-space transformation consistently improves robustness under distribution shift. In particular, *TMA + Filtered Augmentations* achieves the best performance among our variants, improving the OOD average by **4.32%** over

Method	ImageNet	ImageNet-A	ImageNet-V	ImageNet-R	ImageNet-S	Average	OOD Average
CLIP-ViT-B/32	62.22	29.53	54.77	66.23	40.84	50.68	47.84
Existing Back-propagation based TTA approaches							
TPT [39]	63.44	33.11	56.23	69.10	41.09	52.59	49.88
C-TPT [49]	63.85	32.43	56.42	68.45	42.18	52.67	49.87
R-TPT [38]	64.25	36.61	**57.96**	69.94	41.60	54.07	51.53
TTL [21]	64.27	36.15	57.79	**70.86**	42.83	54.38	51.91
TPS [41]	64.25	35.60	57.42	70.05	42.48	53.96	51.39
Ours							
TMA (FA)	64.26	38.01	57.47	70.47	42.70	**54.58** (+3.90)	**52.16** (+4.32)
IMA (FA)	64.13	**38.03**	57.45	70.44	42.54	54.52 (+3.84)	52.12 (+4.28)
TMA (AA)	**64.50**	35.87	57.79	70.19	**43.26**	54.32 (+3.64)	51.78 (+3.94)
IMA (AA)	64.27	36.09	57.78	70.15	42.97	54.25 (+3.57)	51.75 (+3.91)

Table 1. Acc@1 (in %) on Imagenet variants under natural distribution shifts. FA and AA denote Filtered and All Augmentation Strategies.

APPROACH	Caltech101	Pets	Cars	Flowers	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	Average
CLIP-ViT-B/32	91.36	85.06	60.14	64.03	77.37	18.06	62.06	42.97	35.81	61.64	59.85
Existing Back-propagation based TTA approaches											
TPT [39]	92.33	**86.26**	62.22	65.77	78.59	19.47	64.12	43.74	**35.74**	**62.44**	**61.07**
C-TPT [49]	92.05	85.99	60.91	**65.85**	77.72	18.06	63.63	**44.56**	34.20	62.12	60.51
R-TPT [38]	87.02	85.85	52.13	62.53	68.39	19.65	61.33	40.78	28.86	58.00	56.45
TTL [21]	91.85	86.21	62.84	63.22	78.42	18.72	64.30	44.27	32.09	61.99	60.39
TPS [41]	**92.49**	85.80	62.79	64.35	**78.60**	19.44	63.92	44.09	35.68	62.20	60.94
Ours											
TMA (FA)	91.93	85.36	61.91	62.65	77.76	20.25	63.72	42.97	33.68	61.67	60.19 (+0.34)
IMA (FA)	91.81	85.45	61.11	62.65	77.58	20.13	63.37	42.67	33.05	61.43	59.92 (+0.08)
TMA (AA)	92.09	85.58	**63.08**	62.61	77.90	20.58	**64.42**	43.68	32.96	61.54	<u>60.45</u> (+0.60)
IMA (AA)	91.89	85.47	61.85	62.65	77.59	**20.61**	64.30	43.20	31.51	61.35	60.04 (+0.19)

Table 2. Acc@1 (in %) on cross-dataset benchmarks. FA and AA denote Filtered and All Augmentation Strategies.

zero-shot CLIP and surpassing TPT [39] by **2.28%**. The method also surpasses all episodic TTA baselines despite using a significantly lighter adaptation mechanism.

4.5.2. Cross-Dataset Generalisation

Table 2 reports cross-dataset generalization results. Among our variants, TMA + All Augmentations achieves the best average accuracy (underlined) , while maintaining lightweight embedding-space adaptation. All variants consistently improve over zero-shot CLIP. Filtered augmentations perform slightly better on ImageNet-OOD, suggesting that removing unreliable views is beneficial under severe distribution shifts. In contrast, retaining all augmentations improves performance on fine-grained datasets by capturing subtle semantic variations. The small gap between the two indicates that the learned transformation is the primary source of improvement, with the aggregation strategy providing a secondary, task-dependent refinement. Additional results of episodic TTA methods are in Section 8 of the supplementary.

4.6. Ablation Studies

4.6.1. Compute and Memory Analysis

Table 3 compares adaptation time and GPU memory usage of our method with representative TTA approaches operating at different levels: input prompt space (TPT), encoder parameters (TTL), and embedding space (TPS). Our method

Approach	Adaptation Time (s / 1000 samples)	Memory (GB)
TPT [39]	1310.62 ± 1.55	26.11
TTL [21]	111.09 ± 1.61	2.59
TPS [41]	69.66 ± 0.72	0.78
Ours		
IMA (FA)	$\mathbf{68.92 \pm 1.14}$	**0.72**
TMA (FA)	69.93 ± 0.82	**0.72**

Table 3. Average Adaptation Time and Memory Consumption of different approaches on 1000 ImageNet test samples with CLIP ViT-B/32. Note that FA stands for 'Filtered Augmentations.'

Approach	Vanilla Prompt	CoOp	Ensemble
CLIP-ViT-B/32	47.84	48.395	49.795
Ours			
TMA (FA)	**52.16** (+4.32)	**52.41** (+4.02)	**54.05** (+4.26)
IMA (FA)	52.12 (+4.28)	52.36 (+3.97)	54.02(+4.23)
TMA (AA)	51.78 (+3.94)	51.84 (+3.45)	53.55 (+3.75)
IMA (AA)	51.75 (+3.91)	51.79 (+3.40)	53.48 (+3.69)

Table 4. Effect of different textual prototype initializations. We report Acc@1 (in %) of ours comparing with zero-shot CLIP. FA and AA denote 'Filtered' and 'All Augmentations', respectively.

achieves the lowest memory footprint while maintaining highly efficient adaptation time, slightly outperforming TPS. Importantly, unlike methods whose parameter size scales with the number of target classes, our approach uses a fixed number of learnable parameters, making it particularly advantageous in large-class settings. Similar trends are observed for other variants of our framework, highlighting the efficiency of lightweight feature-space realignment compared to prompt- or encoder-based adaptation.

4.6.2. Effect of Different Textual Prototypes

To study the impact of textual prototype quality, we evaluate all four variants of our method on OOD benchmarks of ImageNet using different prompt initialization strategies: the default vanilla prompt, **CoOp** [58] learned prompts, and **prompt ensembling** with 80 handcrafted templates from CLIP [34]. Table 4 shows that stronger textual prototypes consistently improve performance. CoOp yields modest but stable gains, while prompt ensembling provides the largest improvements (e.g., TMA+Filtered Augmentations: $52.16\% \rightarrow 54.05\%$). Importantly, the improvements over zero-shot CLIP remain significant across all prompt initializations, indicating that the primary gains stem from our linear adaptation, with stronger textual prototypes providing an additional orthogonal boost.

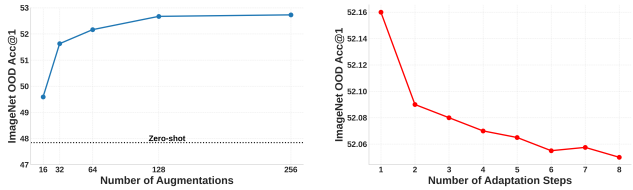


Figure 4. Impact of number of augmented views and optimization steps on mean adaptation performance across Imagenet-OOD datasets using CLIP ViT-B/32 backbone as image encoder.

4.6.3. Effect of Augmentations and Adaptation Steps

We analyze the effect of augmentation count and adaptation steps in Fig. 4. Although our main experiments use 64 augmentations, the method already outperforms zero-shot CLIP by over 1.75% on ImageNet-OOD with only 16 views. Performance improves consistently with more augmentations, suggesting that additional views provide more reliable statistics for entropy minimization and feature realignment. In contrast, increasing the number of optimization steps does not yield noticeable gains, a single step already achieves the best accuracy while avoiding unnecessary computation. Hence, we adopt one-step adaptation as our default setting.

5. Conclusion

We present a test-time adaptation framework that uses lightweight linear transformations to mitigate domain-induced degradation in CLIP-based classification. Rather than tuning pretrained encoders or adding prompt parameters, our method performs instance-level feature-space realignment via entropy-guided optimization. This enables adaptation without altering the pretrained model, preserving its semantic structure while correcting cross-modal misalignment under distribution shifts. A key edge is that the number of learnable parameters depends only on the embedding dimensionality and is independent of the number of target classes, enabling natural scalability to large-vocabulary settings. Despite its simplicity, the approach consistently improves over zero-shot CLIP and remains competitive with prompt- and encoder-based TTA methods across both natural distribution shifts and cross-dataset benchmarks, while requiring substantially lower computational and memory overhead. These findings suggest that minimal, geometry-aware feature correction in the shared embedding space can effectively bridge domain gap, providing a cost-efficient alternative to parameter-heavy adaptation strategies. We hope this work encourages further exploration of lightweight feature-space adaptation for VLMs in resource-constrained and dynamically shifting test environments.

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

[2] Wenxuan Bao, Ruxi Deng, and Jingrui He. Mint: A simple test-time adaptation of vision-language models against common corruptions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2

[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. 6, 1

[4] Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8344–8353, 2022. 1

[5] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 295–305, 2022. 1

[6] 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. 6, 1

[7] 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. 6, 1

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2, 6, 5

[9] Lisa Dunlap, Clara Mohri, Devin Guillory, Han Zhang, Trevor Darrell, Joseph E Gonzalez, Aditi Raghunathan, and Anna Rohrbach. Using language to extend to unseen domains. In *The eleventh international conference on learning representations*, 2022. 5

[10] Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3

[11] 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. 6, 1

[12] 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

[13] Zenghao Guan, Zhou Yucan, Wu Liu, and Xiaoyan Gu. Statistics caching test-time adaptation for vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2

[14] 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. 6, 1

[15] Louis Hémadou, Héléna Vorobieva, Ewa Kijak, and Frédéric Jurie. Adapting without seeing: Text-aided domain adaptation for adapting clip-like models to novel domains. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 5

[16] Dan Hendrycks*, Norman Mu*, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple method to improve robustness and uncertainty under data shift. In *International Conference on Learning Representations*, 2020. 4, 6

[17] 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. 6, 1

[18] 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. 6, 1, 5

[19] 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*, pages 2790–2799. PMLR, 2019. 1

[20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1 (2):3, 2022. 2

[21] Raza Imam, Hanan Gani, Muhammad Huzaifa, and Karthik Nandakumar. Test-time low rank adaptation via confidence maximization for zero-shot generalization of vision-language models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5449–5459. IEEE, 2025. 2, 6, 7, 8, 3, 4

[22] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pages 709–727. Springer, 2022. 1

[23] 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*, pages 14162–14171, 2024. 2

[24] 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

[25] 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. 6, 1

[26] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 3045–3059, 2021. 1

[27] 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*, pages 12888–12900. PMLR, 2022. 1

[28] 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. 1

[29] Yingnan Liu, Rui Qiao, Mong-Li Lee, and Wynne Hsu. Test-time adaptation by causal trimming. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6

[31] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6, 1

[32] 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. 6, 1

[33] 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. 6, 1

[34] 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, 3, 6, 8

[35] 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. 6, 1

[36] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 1

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

[38] Lijun Sheng, Jian Liang, Zilei Wang, and Ran He. R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29958–29967, 2025. 2, 6, 7, 3, 4

[39] 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. 2, 4, 6, 7, 8, 3

[40] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 6, 1

[41] Elaine Sui, Xiaohan Wang, and Serena Yeung-Levy. Just shift it: Test-time prototype shifting for zero-shot generalization with vision-language models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 825–835. IEEE, 2025. 2, 6, 7, 8, 3, 4

[42] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, 2016. 5

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 6

[44] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. 2

[45] 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. 6, 1

[46] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36:61501–61513, 2023. 1

[47] Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, 119(1):3–22, 2016. 6, 1

[48] Zehao Xiao and Cees GM Snoek. Beyond model adaptation at test time: A survey. *arXiv preprint arXiv:2411.03687*, 2024. 1

[49] Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark A. Hasegawa-Johnson, Yingzhen Li, and Chang D. Yoo. C-TPT: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 6, 7, 3, 4

[50] Maxime Zanella and Ismail Ben Ayed. On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23783–23793, 2024. 3

[51] 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

[52] Ce Zhang, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Dual prototype evolving for test-time generalization of vision-language models. *Advances in Neural Information Processing Systems*, 37:32111–32136, 2024. 2

[53] 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, 2024. 1

[54] 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. 2

[55] Yabin Zhang, Wenjie Zhu, Hui Tang, Zhiyuan Ma, Kaiyang Zhou, and Lei Zhang. Dual memory networks: A versatile adaptation approach for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 28718–28728, 2024. 2

[56] Shuai Zhao, Xiaohan Wang, Linchao Zhu, and Yi Yang. Test-time adaptation with CLIP reward for zero-shot generalization in vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[57] 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, 3

[58] 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, 3, 8

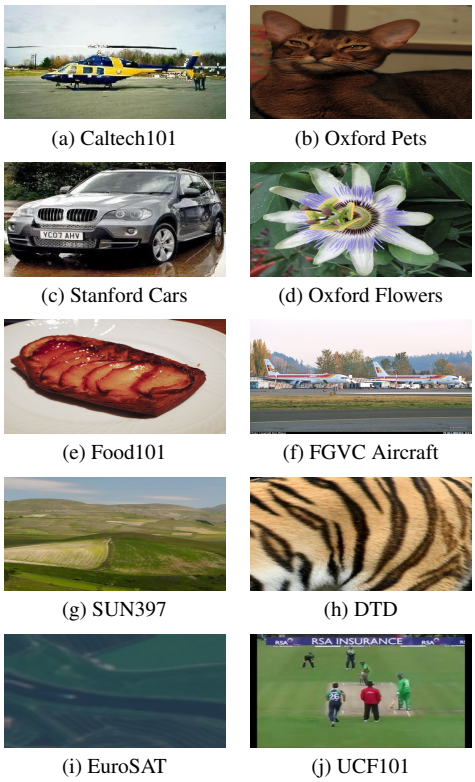
[59] Xingyu Zhu, Shuo Wang, Beier Zhu, Miaoge Li, Yunfan Li, Junfeng Fang, Zhicai Wang, Dongsheng Wang, and Hanwang Zhang. Dynamic multimodal prototype learning in vision-language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2501–2511, 2025. 2

Supplementary Overview

Contents

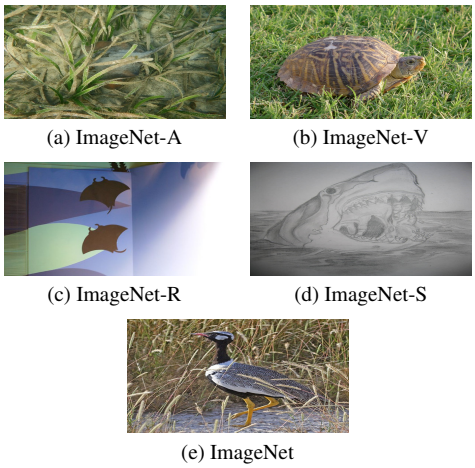
Supplementary Material

6. Overview of Benchmark Details



(a) Caltech101 (b) Oxford Pets (c) Stanford Cars (d) Oxford Flowers (e) Food101 (f) FGVC Aircraft (g) SUN397 (h) DTD (i) EuroSAT (j) UCF101

Figure 5. Visualization on Fine-grained classification datasets.



(a) ImageNet-A (b) ImageNet-V (c) ImageNet-R (d) ImageNet-S (e) ImageNet

Figure 6. Visualization on ImageNet and OOD variants.

6.1. Fine-grained Classification Datasets

To further evaluate cross-dataset generalization, we conduct experiments on ten diverse and publicly available image classification benchmarks spanning a wide range of visual domains. These include fine-grained recognition tasks such as Flowers102 [32] (102 classes, 2,463 test images) and OxfordPets [33] (37 classes, 3,669 test images), transportation categories including StanfordCars [25] (196 classes, 8,041 test images) and FGVC-Aircraft [31] (100 classes, 3,333 test images), and scene understanding with SUN397 [47] (397 classes, 19,850 test images). We also evaluate texture recognition on DTD [6] (47 classes, 1,692 test images), food classification on Food101 [3] (101 classes, 30,300 test images), human action recognition on UCF101 [40] (101 classes, 3,783 test images), satellite imagery classification on EuroSAT [14] (10 classes, 8,100 test images), and general object categorization using Caltech101 [11] (100 classes, 2,465 test images). Together, these datasets vary substantially in granularity, visual complexity, and semantic structure, providing a comprehensive testbed for evaluating the ability of our approach to adapt across heterogeneous domains.

6.2. ImageNet and its OOD variants

We assess robustness to natural distribution shifts using four ImageNet variants that are commonly treated as out-of-distribution (OOD) benchmarks with respect to the original ImageNet [7] dataset. The standard ImageNet validation set contains 1,000 classes and 50,000 test images. These OOD benchmarks provide a standardized and realistic setting for measuring performance degradation under domain changes.

ImageNet-V2 [35] contains 1,000 classes and 10,000 test images and is an independently curated dataset collected from sources distinct from the original ImageNet distribution to capture natural distribution shifts. ImageNet-A [18] consists of 200 classes and 7,500 naturally occurring adversarial images, designed to expose model failures under challenging yet realistic visual conditions. ImageNet-R [17] includes 200 classes and approximately 30,000 artistic and non-photographic renditions of ImageNet categories, such as paintings and sketches, evaluating robustness to significant appearance variations. Finally, ImageNet-Sketch [45] contains 1,000 classes and 50,889 black-and-white sketch images, testing on shape-driven representations.

7. Detailed Algorithmic Descriptions

In this section, we provide detailed algorithmic descriptions of the proposed Image Matrix Adapter (IMA) and Text Matrix Adapter (TMA) methods. The procedures outline the key computational steps involved in performing embedding-space transformations during test-time adaptation.

Algorithm 1: Image Matrix Adapter (IMA)

Input: Input sample x_0
Pre-trained frozen image encoder f_v
Pre-computed textual prototypes $\{t_{c_i} \in R^d\}_{i=1}^K$
Set of augmentations $\mathcal{A}$
Number of additional augmentations $(N-1)$
Aggregation Strategy mode (Filtered/All)
AdamW optimizer Opt

Output: Predicted class label c_i from the K classes

function ADAPT $(x_0, f_v, \{t_{c_i}\}_{i=1}^K, \mathcal{A}, N, \text{Opt}, \text{mode})$
  Sample $x_1, x_2, ..., x_{N-1} \in \mathcal{U}(\mathcal{A})$
  $v_j = f_v(x_j) \in R^d$, for j = 0 to N - 1
  Compute $p(c_i \mid (x_j, t_{c_i})) = \dfrac{\exp((v_j^T t_{c_i})/\tau)}{\sum_{k=1}^K \exp((v_j^T t_{c_k})/\tau)}$ for $j \in \{0, 1, \ldots, N-1\}$
  $H(x_j) = -\sum_{i=1}^K p(c_i \mid (x_j, t_{c_i})) \log(p(c_i \mid (x_j, t_{c_i})))$ for $j \in \{0, 1, \ldots, N-1\}$
  Initialize $W_v \leftarrow \mathbf{I}_{\mathbf{d}\times\mathbf{d}}$
  if mode $=$ *Filtered* **then**
    $S = \{x_j : \mathbb{1}\{H(x_j) \le \tau\}\}$ where,
    $\tau = \text{Percentile}_\rho(\{H(x_j)\}_{j=0}^{N-1})$
    The set $\{x_j'\}_{j=0}^{s-1} \equiv S$ denote Filtered Augmentations
    $v_j' = f_v(x_j')$ for $j \in \{0, 1, \ldots, s-1\}$
    Compute
    $\tilde{p}(c_i \mid (x_0, W_v), t_{c_i}) = \dfrac{1}{s}\sum_{j=0}^{s-1} p(c_i \mid (x_j', W_v), t_{c_i})$ where
    $p(c_i \mid (x_j', W_v), t_{c_i}) = \dfrac{\exp(sim((W_v v_j'), t_{c_i})/\tau)}{\sum_{k=1}^K \exp(sim((W_v v_j'), t_{c_k})/\tau)}$
    Compute $\mathcal{L}$ by eq 3 using $\tilde{p}$
  end
  else if mode $=$ *All* **then**
    Compute α_j by eq 13, for $j \in \{0, 1, \ldots, N-1\}$
    Compute $\mathcal{L}$ by eq 11, where $W = W_v$
  end
  Compute $\partial\mathcal{L}$
  Update $W_v := W_v \ - \text{Opt}(\partial\mathcal{L})$
  Return W_v
end

function INFERENCE $(x_0, f_v, \{t_{c_i}\}_{i=1}^K, \mathcal{A}, \mathcal{N}, \text{Opt}, \text{mode})$
  $v_0 = f_v(x_0)$
  $W_v^* = \text{ADAPT}(x_0, f_v, \{t_{c_i}\}_{i=1}^K, \mathcal{A}, \mathcal{N}, \text{Opt}, \text{mode})$
  $v_0^{\text{new}} = \frac{W_v^* v_0}{\|W_v^* v_0\|_2}$
  Compute $p(c_i \mid (x_0, W_v^*), t_{c_i}) = \dfrac{\exp((v_0^{\text{new}})^T t_{c_i})}{\sum_{k=1}^K \exp((v_0^{\text{new}})^T t_{c_k})}$ for $i \in \{1, 2, \ldots, K\}$
  Return $\arg\max_{c_i} p(c_i \mid (x_0, W_v^*), t_{c_i})$
end

Algorithm 2: Text Matrix Adapter (TMA)

Input: Input sample x_0
Pre-trained frozen image encoder f_v
Pre-computed textual prototypes $\{t_{c_i} \in R^d\}_{i=1}^K$
Set of augmentations $\mathcal{A}$
Number of additional augmentations $(N-1)$
Aggregation Strategy mode (Filtered/All)
AdamW optimizer Opt

Output: Predicted class label c_i from the K classes

function ADAPT $(x_0, f_v, \{t_{c_i}\}_{i=1}^K, \mathcal{A}, N, \text{Opt}, \text{mode})$
  Sample $x_1, x_2, ..., x_{N-1} \in \mathcal{U}(\mathcal{A})$
  $v_j = f_v(x_j) \in R^d$, for j = 0 to N - 1
  Compute $p(c_i \mid (x_j, t_{c_i})) = \dfrac{\exp((v_j^T t_{c_i})/\tau)}{\sum_{k=1}^K \exp((v_j^T t_{c_k})/\tau)}$ for $j \in \{0, 1, \ldots, N-1\}$
  $H(x_j) = -\sum_{i=1}^K p(c_i \mid (x_j, t_{c_i})) \log(p(c_i \mid (x_j, t_{c_i})))$ for $j \in \{0, 1, \ldots, N-1\}$
  Initialize $W_t \leftarrow \mathbf{I}_{\mathbf{d}\times\mathbf{d}}$
  if mode $=$ *Filtered* **then**
    $S = \{x_j : \mathbb{1}\{H(x_j) \le \tau\}\}$ where,
    $\tau = \text{Percentile}_\rho(\{H(x_j)\}_{j=0}^{N-1})$
    The set $\{x_j'\}_{j=0}^{s-1} \equiv S$ denote Filtered Augmentations
    $v_j' = f_v(x_j')$ for $j \in \{0, 1, \ldots, s-1\}$
    Compute
    $\tilde{p}(c_i \mid x_0, (t_{c_i}, W_t)) = \dfrac{1}{s}\sum_{j=0}^{s-1} p(c_i \mid x_j', (t_{c_i}, W_t))$ where
    $p(c_i \mid x_j', (t_{c_i}, W_t)) = \dfrac{\exp(sim(v_j', (W_t t_{c_i}))/\tau)}{\sum_{k=1}^K \exp(sim(v_j', (W_t t_{c_k}))/\tau))}$
    $\mathcal{L} = -\sum_{i=1}^K \tilde{p}(c_i \mid x_0, (t_{c_i}, W_t)) \log(\tilde{p}(c_i \mid x_0, (t_{c_i}, W_t)))$
  end
  else if mode $=$ *All* **then**
    Compute α_j by eq 13 (in main), for $j \in \{0, 1, \ldots, N-1\}$
    Compute $\mathcal{L}$ by eq 11 (in main), where $W = W_t$
  end
  Compute $\partial\mathcal{L}$
  Update $W_t := W_t \ - \text{Opt}(\partial\mathcal{L})$
  Return W_t
end

function INFERENCE $(x_0, f_v, \{t_{c_i}\}_{i=1}^K, \mathcal{A}, \mathcal{N}, \text{Opt}, \text{mode})$
  $v_0 = f_v(x_0)$
  $W_t^* = \text{ADAPT}(x_0, f_v, \{t_{c_i}\}_{i=1}^K, \mathcal{A}, \mathcal{N}, \text{Opt}, \text{mode})$
  $t_{c_i}^{\text{new}} = \frac{W_t^* t_{c_i}}{\|W_t^* t_{c_i}\|_2}$, for $i \in \{1, 2, \ldots, K\}$
  Compute $p(c_i \mid x_0, (t_{c_i}, W_t^*)) = \dfrac{\exp((v_0^T t_{c_i}^{new})}{\sum_{k=1}^K \exp((v_0^T t_{c_k}^{new})}$ for $i \in \{1, 2, \ldots, K\}$
  Return $\arg\max_{c_i} p(c_i \mid x_0, (t_{c_i}, W_t^*))$
end

8. Additional Experimental Results

This section contains results other episodic approaches such as ZERO, MTA and RLCF using our default setting (ViT-B/32 backbone). Also, results are reported using baselines in Section 4 of main paper with CLIP ViT-B/16 and RN50.

APPROACH	TPT	TTL	TPS	ZERO [10]	MTA [50]	RLCF [56]	TMA (FA)	IMA (FA)	TMA (AA)	IMA (AA)
ImageNet [39]	63.44	64.27	64.25	**65.31**	64.79	63.64	64.26	64.13	64.50	64.27
ImageNet OOD [49]	49.88	51.91	51.39	50.88	51.60	51.14	**52.16**	52.12	51.78	51.75

Table 5. Results on cross-dataset benchmarks with ViT-B/32 backbone. FA and AA denote Filtered and All Augmentation Strategies. We report Acc@1 (in %). The best result across all methods is shown in bold.

APPROACH	Caltech101	Pets	Cars	Flowers	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	Average
CLIP-**ViT-B/16**	93.31	88.25	65.33	67.40	83.64	23.91	63.05	44.39	42.22	65.24	63.68
Existing Back-propagation based TTA approaches											
TPT [39]	**94.04**	87.71	66.48	**69.47**	**84.46**	24.00	**65.19**	**46.04**	42.37	67.27	**64.70**
C-TPT [49]	93.63	**88.83**	65.96	69.18	83.92	24.03	64.42	45.45	39.65	66.03	64.11
R-TPT [38]	89.70	86.81	62.87	67.19	80.46	24.39	63.76	42.85	32.37	61.93	61.23
TTL [21]	93.83	87.74	**66.78**	67.24	84.08	**25.11**	65.14	45.15	**42.73**	**67.51**	64.53
TPS [41]	94.00	87.14	67.03	67.64	84.25	24.27	64.64	45.69	43.54	66.69	64.49
Ours											
TMA (FA)	93.47	86.35	66.26	66.75	83.20	24.60	64.77	44.62	40.56	66.27	<u>63.69</u>
IMA (FA)	93.35	86.29	65.69	66.50	82.84	24.54	64.53	44.62	39.58	66.27	63.42
TMA (AA)	93.43	87.08	66.47	66.54	83.55	**25.11**	65.17	44.27	38.01	66.40	63.60
IMA (AA)	93.23	86.86	65.86	66.30	83.26	24.75	65.17	44.03	36.05	66.48	63.20

Table 6. Results on cross-dataset benchmarks with ViT-B/16 backbone. FA and AA denote Filtered and All Augmentation Strategies. We report Acc@1 (in %). The best result across all methods is shown in bold, while the best result of our method is underlined.

APPROACH	Caltech101	Pets	Cars	Flowers	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	Average
CLIP-**RN50**	85.68	83.62	55.29	61.67	73.96	15.69	59.25	40.43	23.69	58.90	55.81
Existing Back-propagation based TTA approaches											
TPT [39]	86.82	**84.71**	57.11	**62.48**	**74.93**	16.20	**60.87**	**41.55**	24.00	**60.64**	**56.93**
C-TPT [49]	86.61	83.43	55.47	62.44	74.29	16.68	60.56	41.25	23.19	59.34	56.33
R-TPT [38]	71.08	82.99	50.54	59.40	66.95	15.96	55.67	37.71	19.69	50.30	51.03
TPS [41]	**87.22**	84.19	57.16	61.75	74.54	17.22	60.42	40.84	**26.00**	59.82	56.92
Ours											
TMA (FA)	86.41	83.40	57.28	59.44	71.89	16.62	59.92	40.19	23.17	58.97	<u>55.73</u>
IMA (FA)	86.21	83.43	56.87	59.07	71.60	15.87	59.50	39.30	22.60	58.79	55.32
TMA (AA)	85.35	83.97	**57.82**	59.72	72.32	**17.79**	60.49	40.96	16.07	59.37	55.39
IMA (AA)	85.07	83.81	57.18	59.32	71.80	17.10	60.16	40.48	15.74	59.00	54.97

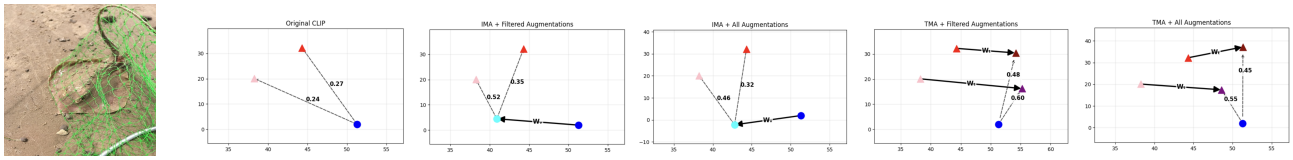
Table 7. Results on cross-dataset benchmarks with ResNet-50 backbone. FA and AA denote Filtered and All Augmentation Strategies. We report Acc@1 (in %). The best result across all methods is shown in bold, while the best result of our method is underlined.

Method	ImageNet-A	ImageNet-V	ImageNet-R	ImageNet-S	OOD Average
CLIP-**ViT-B/16**	47.80	60.84	73.99	46.15	57.20
Existing Back-propagation based TTA approaches					
TPT [39]	52.89	62.57	76.92	47.36	59.94
C-TPT [49]	50.55	62.43	75.67	47.19	58.96
R-TPT [38]	47.71	60.59	74.14	42.75	56.30
TTL [21]	55.21	62.97	**77.24**	47.56	**60.75**
TPS [41]	55.23	62.88	76.61	**47.66**	60.60
Ours					
TMA (FA)	56.47	62.77	76.53	47.24	**60.75**
IMA (FA)	**56.55**	62.62	76.46	47.19	60.71
TMA (AA)	54.57	**63.10**	76.22	47.51	60.35
IMA (AA)	54.41	63.04	76.22	47.32	60.25

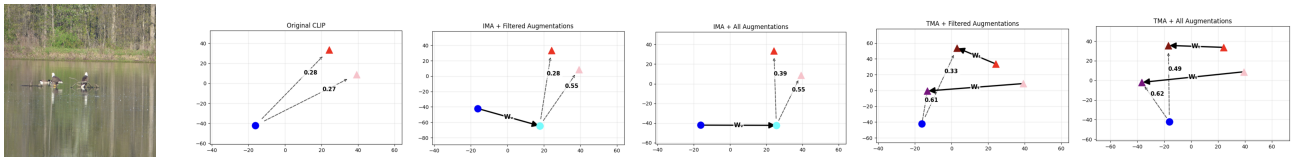
Table 8. Results on ImageNet-OOD datasets using the ViT-B/16 backbone. We report Acc@1 (in %). FA and AA denote the Filtered and All augmentation strategies, respectively. The best result across all methods is shown in bold.

Method	ImageNet-A	ImageNet-V	ImageNet-R	ImageNet-S	OOD Average
CLIP-**RN50**	21.84	51.52	56.09	33.34	40.70
Existing Back-propagation based TTA approaches					
TPT [39]	25.54	52.61	**58.93**	**35.17**	43.06
C-TPT [49]	23.96	**54.14**	56.67	34.68	42.36
R-TPT [38]	24.35	54.12	57.73	33.87	42.52
TPS [41]	**27.18**	52.84	57.34	34.92	**43.07**
Ours					
TMA (FA)	26.40	53.14	57.25	33.56	42.64
IMA (FA)	26.29	52.68	56.78	32.91	42.17
TMA (AA)	25.39	53.36	57.54	34.28	<u>42.65</u>
IMA (AA)	25.65	52.97	56.93	34.00	42.39

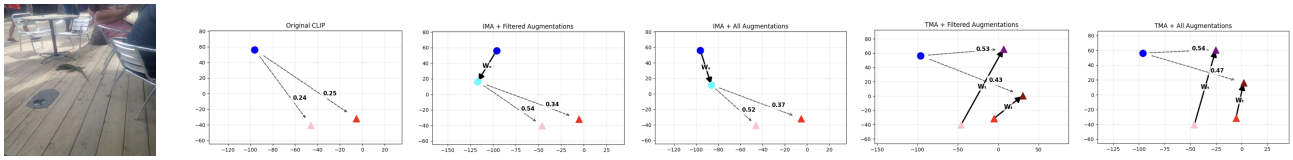
Table 9. Results on ImageNet-OOD datasets using the RN50 backbone. We report Acc@1 (in %). FA and AA denote the Filtered and All augmentation strategies, respectively. The best result across all methods is shown in bold, while the best result of our method is underlined.



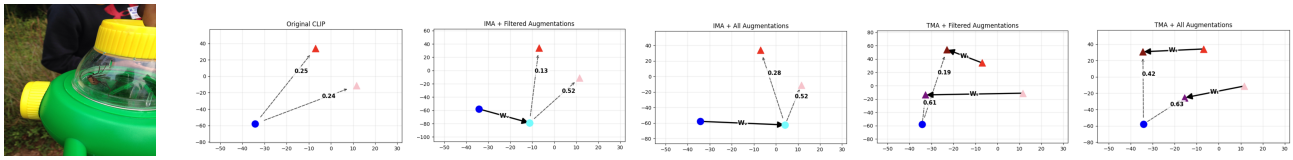
Ground Truth : **stingray**; CLIP Zero-Shot Prediction : hermit crab



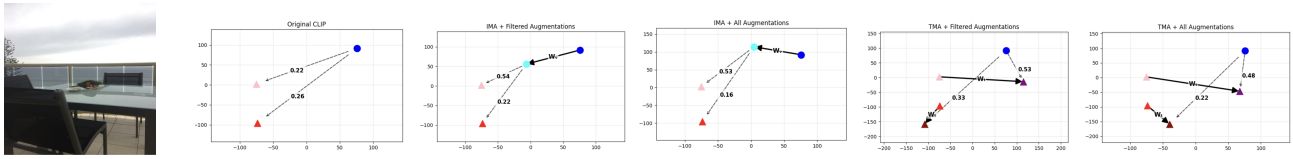
Ground Truth : **bald eagle**; CLIP Zero-Shot Prediction : goose



Ground Truth : **green iguana**; CLIP Zero-Shot Prediction : american bullfrog



Ground Truth : **harvestman**; CLIP Zero-Shot Prediction : rhinoceros beetle



Ground Truth : **lorikeet**; CLIP Zero-Shot Prediction : wine bottle

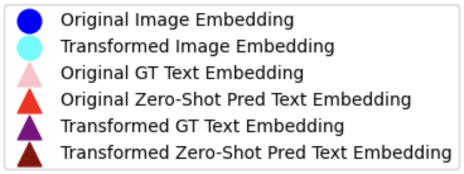


Figure 7. Visualization of embedding transformations under different adaptation strategies. In each row, the leftmost image corresponds to the raw test sample from the ImageNet-A [18] dataset. All subsequent plots visualize the corresponding image embeddings and textual prototypes projected onto a 2D space using t-SNE for interpretability. The second plot shows the original zero-shot embedding configuration of CLIP ViT-B/32 [8], where it predicts incorrectly i.e., it is giving lesser similarity score to ground truth class label. The remaining four plots illustrate the effect of the proposed linear adaptation strategies: IMA with Filtered Augmentations (FA), IMA with All Augmentations (AA), TMA with Filtered Augmentations (FA), and TMA with All Augmentations (AA). Through entropy minimization, these adaptations adjust the image or text embeddings to increase the confidence of the ground-truth label, effectively making it **argmax** in the final inference aligning the model prediction with the correct class. The markers of legend denote image and text embeddings. Dashed arrows denote cosine similarity between embeddings, while bold arrows signify the direction of embedding transformations in the projected space.