arXiv:2608.18339v1 [cs.CV] 18 Aug 2026

From Inference to Adaptation: A Unified Optimal Transport View of Vision Language Model

Qi Yu[1]* **Zhichen Zeng**[1] **Katherine Tieu**[1] **Xiyuan Yang**[1] **Ruizhong Qiu**[1] **Yuchen Yan**[2]
Lihui Liu[3] **Yanjun Zhao**[1] **Lingjie Chen**[1] **Jingrui He**[1] **Hanghang Tong**[1]

[1] University of Illinois Urbana-Champaign [2] Amazon [3] Wayne State University

Abstract

Vision-language models (VLMs) have demonstrated remarkable zero-shot capabilities yet remain sensitive to real-world distribution shifts during inference. Although significant efforts are devoted to adapting VLMs at test time, they rely heavily on noisy pseudo-labels predicted directly from raw embedding similarities during inference, which are unreliable under distribution shift and mislead the adaptation. To avoid noise amplification, existing works craft coarse-grained surrogate objectives during adaptation, which fail to explicitly model sample-level relationships across different modalities, creating objective mismatch with inference, thus leading to marginal performance improvement. In this work, we aim to bridge the detached objectives of inference and adaptation for VLMs, and propose a principled VLM TTA method called ORIGIN. For *VLM inference*, we formulate the zero-shot image classification task as a cross-modal alignment problem encoded via a Wasserstein OT formulation, providing robust pseudo-labels at the sample-level to effectively adapt VLMs. For *VLM adaptation*, we adopt a soft-label InfoNCE loss to adapt VLMs based on the OT-induced pseudo-labels, leveraging fine-grained supervisions to explicitly model relationships of individual image-text pairs via contrastive learning, which empowers accurate inference at the same granularity. Moreover, we theoretically reveal that the InfoNCE loss can be neatly reformulated as a Wasserstein OT formulation, thereby unifying the objectives of the inference and adaptation of VLMs to achieve their mutual benefits. Extensive experiments demonstrate the effectiveness and efficiency of our methods, outperforming the best-performing methods by up to 7% with state-of-the-art efficiency.

1 Introduction

Vision-language models (VLMs), such as CLIP [Radford et al., 2021], learn to project image and text embeddings into a shared embedding space, and have achieved remarkable zero-shot capability across diverse multimodal learning tasks, such as image classification [Radford et al., 2021], image captioning [Li et al., 2022], and semantic segmentation [Lüddecke and Ecker, 2022]. Despite the success of pre-trained VLMs, the aligned embedding spaces are vulnerable to real-world distribution shifts, such as image corruption [Hendrycks and Dietterich, 2019], causing significant performance degradation of VLMs at deployment. To mitigate this issue, a predominant approach is to adapt pre-trained VLMs at test time, i.e., test-time adaptation (TTA), based on unlabeled test data. This paradigm allows pre-trained VLMs to adapt dynamically to shifted domains at deployment without accessing large-scale source training data or test labels.

Although existing TTA methods have demonstrated strong improvement across different benchmarks, they bear the following two key limitations which lead to suboptimal adaptation performance. Firstly

*Correspondence to: Qi Yu <qiyu6@illinois.edu>

Preprint.

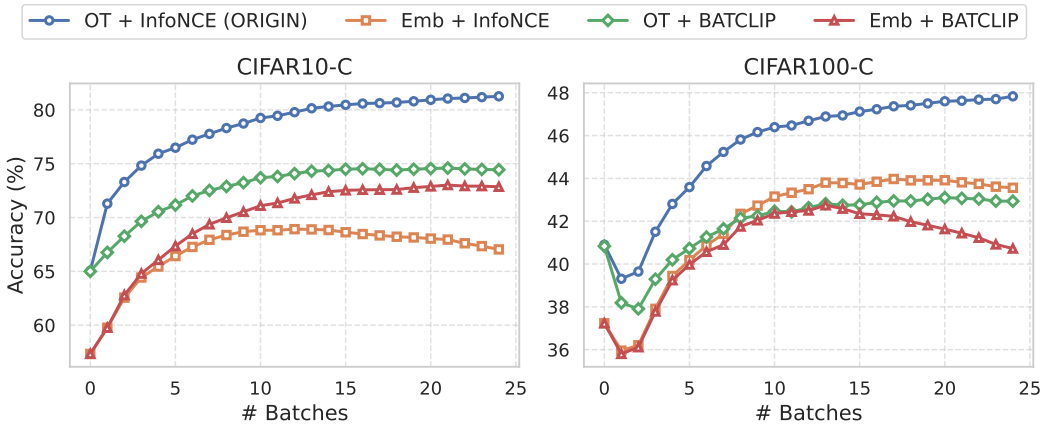

Figure 1: Training dynamics of different combinations of inference and adaptation objective on CIFAR10-C. **(1)** Embedding-based inference are consistently outperformed by OT-based inference under the same loss function, and may even leads to performance degradation during adaptation due to excessive noise amplification; **(2)** ORIGIN adopts InfoNCE loss with OT-induced pseudo-labels to explicitly model sample-level relationships across modalities, increasing accuracy monotonically during adaptation and consistently achieves the best performance.

(Limitation #1), existing TTA methods predominantly rely zero-shot predictions of VLMs inferred from the raw embedding similarity of images and texts under distribution shifts, thus generating noisy and unreliable pseudo-labels that later limit or even misguide the adaptation of pre-trained VLMs [Shu et al., 2022a, Li et al., 2024a, Sheng et al., 2025]. As shown in Table 2 and Figure 1, pseudo-labels inferred from raw embedding similarities suffer from poor accuracy, leading to suboptimal or even degraded performance during adaptation.

Secondly *(Limitation #2)*, limited by the noisiness of pseudo-labels, existing TTA methods rely heavily on crafting surrogate objectives on aggregated class-level prototypes (e.g., average embedding of a class) [Maharana et al., 2025, Bao et al., 2025] to avoid excessive noise amplification. However, class-level supervisions are inherently coarse-grained, treating individual samples uniformly. Therefore, they fail to model cross-modal relationships between individual image-text pairs, which inherently limits the adaptation performance evaluated on sample-level alignment tasks during inference. As shown in Figure 1 and Table 2, while the objective of BATCLIP [Maharana et al., 2025], the best-performing baseline, can sometimes avoid noise amplification during adaptation, its performance are inherently limited by the class-level design of its adaptation objective.

In this paper, we propose a principled VLM TTA framework which leverages optimal transport (OT) to bridge the inference and adaptation of VLMs at test time, called ORIGIN. At the *inference* stage, we formulate the zero-shot image classification task as an alignment problem across the image and text modalities, whose results are inferred from the transport plan of a Wasserstein OT formulation. Empowered by the global view and constrained nature of OT [Chen et al., 2020, Yu et al., 2025b, Zhu et al., 2025, Yu et al., 2025a], predictions inferred from the solved transport map is significantly more accurate than that inferred from raw embedding similarities, as shown in Table 2, therefore generating more robust and reliable pseudo-labels for TTA *(Limitation 1)*. At the *adaptation* stage, equipped with the reliable pseudo-labels derived from the Wasserstein OT formulation, we directly adopt the soft-label InfoNCE loss, aligning with the pre-training objective of VLMs, to explicitly model the embedding relationships of image-text pairs via contrastive learning, leverging fine-grained supervision signals to unleash the full potential of test data for adapting the model at test time *(Limitation 2)*. As shown in Figure 1, ORIGIN adopts the soft-label InfoNCE loss with OT-induced pseudo-labels, guiding the adaptation effectively and leading to consistent improvement in accuracy.

Finally, to further bridge the test-time inference and adaptation of VLMs, we theoretically prove that the InfoNCE loss can be reformulated to an entropic Wasserstein OT formulation. In this way, the inference and adaptation of VLMs are no longer decoupled stages but a mutually beneficial process with a shared objective function through the lens of OT. While robust and reliable pseudo-labels produced at the inference stage help improve the adaptation of VLMs, effective adaptation of VLMs also aligns disparate embeddings of text and image modalities under distribution shifts, thus improving

the inference performance. By alternating between optimizations of the OT transport plan and the parameters of VLMs, we achieve the mutual benefits of robust inference and fine-grained adaptation, which maximally improve the TTA performance of VLMs. To evaluate the effectiveness of our proposed ORIGIN, we conduct extensive experiments on three standard benchmarking datasets, showing that our method significantly outperforms state-of-the-art TTA methods with up to 7% mean accuracy improvement without additional latency. Our main contribution are summarized as follows:

- **Unified View.** To our best knowledge, we are the first to theoretically bridge the inference and adapation of VLMs at test-time through the lens of optimal transport.
- **Novel Methodology.** We proposed a novel VLM TTA methods based on optimal transport which aligns the loss functions of adaptation with the pretraining objective of VLMs.
- **Strong Performance.** Extensive experiments on three standard VLM TTA benchmarks demonstrate significant performance gain compared to state-of-the-art baselines.

2 Preliminaries

In this section, we first introduce preliminaries on VLM TTA in Section 2.1, followed by optimal transport in Section 2.2. In this paper, we use bold uppercase letters for matrices (e.g., $\mathbf{S}$), bold lowercase letters for vectors (e.g., $\boldsymbol{\mu}$), calligraphic uppercase letters for sets (e.g., $\mathcal{T}$) and lowercase letters for scalars (e.g., k).

2.1 Test-time Adaptation of VLMs

We denote the text space by $\mathcal{T}$ and image space by $\mathcal{V}$. A pre-trained VLM consists of a text encoder $f_\phi : \mathcal{T} \to \mathbb{R}^d$ and an image encoder $f_\theta : \mathcal{V} \to \mathbb{R}^d$, which map images and texts into a shared embedding space of dimension d. For a zero-shot image classification task with n different classes, the image encoder f_θ projects an image v to an embedding $\mathbf{v} \in \mathbb{R}^d$, and the text encoder f_ϕ projects text descriptions of c different classes, e.g., "a photo of a <class>" to embeddings $\{\mathbf{t}_j \in \mathbb{R}^d\}_{j=1}^c$. Predictions are inferred by comparing the embedding similarity of the test image and text descriptions of different classes, e.g., $\arg\max_j \mathbf{v}^\top \mathbf{t}_j$. In this paper, we consider the online TTA setting, where test images under distribution shifts arrives sequentially in a batch manner [Maharana et al., 2025].

2.2 Optimal Transport

OT is becoming a popular and effective tool for aligning distributional structures [Santambrogio, 2015]. Formally, let $\boldsymbol{\mu} = \sum_{i=1}^n \mu_i \delta_{x_i}$ and $\boldsymbol{\nu} = \sum_{j=1}^m \nu_i \delta_{y_j}$ be two discrete probability distributions where δ denotes the Dirac measure, the discrete OT problem seeks an optimal transport plan $\mathbf{S}$ that minimizes the total transport cost as follows:

$$\min_{\mathbf{S} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \langle \mathbf{C}, \mathbf{S} \rangle \tag{1}$$

where $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu}) := \left\{ \mathbf{S} \in \mathbb{R}_+^{n \times m} | \mathbf{S}\mathbf{1}_m = \boldsymbol{\mu}, \mathbf{S}^\top \mathbf{1}_n = \boldsymbol{\nu} \right\}$, and $\mathbf{C} \in \mathbb{R}^{n \times m}$ is the cost matrix measuring the cost of transporting mass from different units of supports. The optimal value of Eq. (1) defines the *Wasserstein distance* between $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ under the cost matrix $\mathbf{C}$, and the resulting transport plan encodes the soft correspondence between points from the two distributions.

To solve the Wasserstein OT formulation efficiently, [Peyré and Cuturi, 2019] introduces entropic regularization into Eq. (1) to approximate the original OT formulation:

$$\min_{\mathbf{S} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \langle \mathbf{C}, \mathbf{S} \rangle - \epsilon\, H(\mathbf{S}) \tag{2}$$

where $H(\mathbf{S}) := -\sum_{i,j} \mathbf{S}_{i,j} (\log \mathbf{S}_{i,j} - 1)$ and $\epsilon > 0$ denotes the entropic regularization weight. Eq. (2) yields an ϵ-strongly convex optimization problem solvable via the Sinkhorn algorithm with a quadratic complexity [Nemirovski and Rothblum, 1999]. Specifically, The optimal solution of Eq. (2) yields the following scaling form:

$$\mathbf{S}^* = \mathrm{diag}(\mathbf{a})\, \mathbf{K}\, \mathrm{diag}(\mathbf{b}), \tag{3}$$

where $\mathbf{K} := \exp(-\mathbf{C}/\epsilon)$ is the Gibbs kernel, and $\boldsymbol{\mu} \in \mathbb{R}_+^n$, $\boldsymbol{\nu} \in \mathbb{R}_+^m$ are scaling vectors which enforce the marginal constraints. The Sinkhorn algorithm solves for $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ via iterative matrix scaling:

$$\mathbf{a}^{(t+1)} = \boldsymbol{\mu} \oslash (\mathbf{K}\mathbf{b}^{(t)}), \qquad \mathbf{b}^{(t+1)} = \boldsymbol{\nu} \oslash (\mathbf{K}^\top \mathbf{a}^{(t+1)}), \tag{4}$$

where $\oslash$ denotes element-wise division.

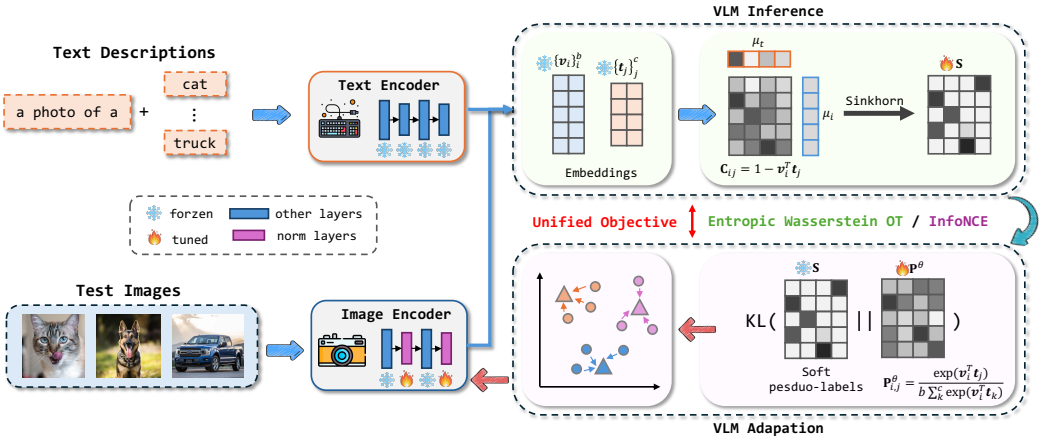

Figure 2: An overview of ORIGIN. VLMs first encode test images via the image encoder and text descriptions of different classes via the text encoder. **During inference**, ORIGIN infers zero-shot image classification by solving an entropic Wasserstein OT problem. **During adaptation**, ORIGIN explicitly models the embedding relationships of shifted modalities at the sample-level via a soft-label InfoNCE loss, and optimizes normalization layers of the image encoder.

3 Methodology

In this section, we present the proposed ORIGIN. We first introduce a Wasserstein OT formulation which predicts image-text alignment robustly during inference, in Section 3.1. Then, we introduce the adopted soft-label InfoNCE during adaptation which explicitly models sample-level relationships via contrastive learning, in Section 3.2. Finally, we theoretically unifies the objectives of ORIGIN at the inference and adaptation stage, in Section 3.3. An overview of ORIGIN is shown in Figure 2.

3.1 OT-based Inference

Pre-trained VLMs are trained to project different modalities, i.e., images and texts, into the same embeddings space, making them directly comparable based on embedding similarities. However, the unified embedding space of VLMs are vulnerable under distribution shifts, making embedding similarities noisy to infer the actual semantic similarity between shifted images and texts. OT, in contrast, leverages a constrained optimization framework to produce noise-reduced similarity across different modalities [Chen et al., 2020]. Therefore, at the inference stage of VLMs, we propose to use the raw embedding distances across shifted modalities as their transport cost at a distribution level, and instead infer image-text alignment by solving an entropic Wasserstein OT formulation in Eq. (5).

Formally, given a batch of test images $\mathcal{V} := \{v_i\}_{i=1}^{b}$ of size b and a set of text descriptions $\mathcal{T} := \{t_j\}_{j=1}^{c}$ of c different text classes, we predict image-text alignment by solving the following entropic Wasserstein OT formulation:

$$\min_{\mathbf{S}\in\Pi(\boldsymbol{\mu}_i,\boldsymbol{\mu}_t)} \langle \mathbf{C}, \mathbf{S}\rangle - \epsilon\, H(\mathbf{S}) \quad \mathbf{C}_{i,j} := 1 - \mathbf{v}_i^\top \mathbf{t}_j \tag{5}$$

where $\mathbf{v}_i = f_\theta(v_i)$, $\mathbf{t}_j = f_\phi(t_j)$. As we will show later in Section 3.3, such cost design provides a unified view of VLM inference and adaption. $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_t$ are two marginal distributions defined over the image batch and text classes respectively. We assume a uniform distribution at the image level, i.e., $\boldsymbol{\mu}_i := \sum_i^b \frac{1}{b}\delta_{v_i}$, treating each test image with equal importance. For text-level distribution $\boldsymbol{\mu}_t$, we assume a uniform distribution before adaptation and dynamically adjust $\boldsymbol{\mu}_t$ according to the predictions of previous test batches by exponential moving average (EMA), i.e.,

$$\boldsymbol{\mu}_t^{(k)} = \alpha\hat{\boldsymbol{\mu}}_t^{(k-1)} + (1-\alpha)\boldsymbol{\mu}_t^{(k-1)} \quad \boldsymbol{\mu}_t^{(0)} := \sum_{i=1}^{c} \frac{1}{c}\delta_{t_i} \tag{6}$$

where $\boldsymbol{\mu}_t^{(k)}$ denotes the text-level marginal distribution for OT, and $\hat{\boldsymbol{\mu}}_t^{(k-1)}$ denotes the predicted text label distribution at the (k-1)-th batch. α is the EMA parameter. For a test image v_i, its predicted label $\tilde{c}_i$ based on image-text alignment is inferred from the solved transport map $\mathbf{S}^*$, i.e., $\tilde{c} = \arg\max_j \mathbf{S}^*_{i,j}$.

Eq. (5) yields an ϵ-strongly convex optimization problem that can be solved more efficiently via the Sinkhorn algorithm [Peyré and Cuturi, 2019], as shown in Eq. (3) and Eq. (4).

Instead of relying on noisy embedding similarities which are unreliable under distribution shifts, ORIGIN infer image-text alignment from a global view by leveraging the constrained nature of OT [Chen et al., 2020, Zhu et al., 2025, Yu et al., 2025b], therefore leading to more accurate predictions which produces robust and reliable pseudo-labels for VLM adaptation.

3.2 Fine-grained Adaptation

Empowered by the robust predictions of test image labels via OT, ORIGIN directly adopts a soft-label InfoNCE loss for image-to-text alignment as the optimization objective during adaptation, modeling cross-modal relationships explicitly at the same sample-level as VLM inference. Formally, for a batch of test image $\mathcal{V} := \{v_i\}_i^b$ and a set of text descriptions $\mathcal{T} := \{t_j\}_j^c$ of c different text classes, ORIGIN optimizes the following soft-label InfoNCE loss in Eq. (7).

$$\min_{\theta} \mathrm{KL}(\mathbf{S} \| \tilde{\mathbf{P}}^{\theta}), \quad \tilde{\mathbf{P}}^{\theta}_{i,j} := \frac{\exp(\mathbf{v}_i^\top \mathbf{t}_j / \tau)}{b \sum_{k=1}^{c} \exp(\mathbf{v}_i^\top \mathbf{t}_k / \tau)} \tag{7}$$

where $\mathrm{KL}(\cdot \| \cdot)$ denotes the KL divergence, $\mathbf{S}$ is the transport map of OT at the inference stage serving as soft pseudo-labels, and τ is a scalar temperature. Recall that the pre-training objective of CLIP consists of two hard-label InfoNCE losses which simultaneously conduct image-to-text and text-to-image alignment in the embedding space, as shown in Eq. (8) [Shi et al., 2024]

$$\min_{\theta} \mathrm{KL}(\mathbf{P} \| \tilde{\mathbf{P}}^{\theta,\phi}) + \mathrm{KL}(\mathbf{P} \| \hat{\mathbf{P}}^{\theta,\phi}) \tag{8}$$

where $\mathbf{P}$ consists of one-hot vectors denoting hard alignment of paired images and texts, and $\tilde{\mathbf{P}}^{\theta,\phi}, \hat{\mathbf{P}}^{\theta,\phi}$ are defined as follows

$$\tilde{\mathbf{P}}^{\theta,\phi}_{i,j} := \frac{\exp(\mathbf{v}_i^\top \mathbf{t}_j / \tau)}{b \sum_{k=1}^{c} \exp(\mathbf{v}_i^\top \mathbf{t}_k / \tau)}, \quad \hat{\mathbf{P}}^{\theta,\phi}_{i,j} := \frac{\exp(\mathbf{v}_i^\top \mathbf{t}_j / \tau)}{c \sum_{k=1}^{b} \exp(\mathbf{v}_k^\top \mathbf{t}_j / \tau)} \tag{9}$$

Comparing Eq. (7) and Eqs. (8)-(9) reveals that the adaptation objective of ORIGIN essentially **1)** replaces ground-truth hard label of $\mathbf{P}$ with OT-induced soft pseduo-label $\mathbf{S}$ since test labels are unavailable during inference at test-time, and **2)** solely adapts image encoder and aligns shifted image embeddings to text embeddings since distribution shifts typically happens on the image side [Bao et al., 2025, Maharana et al., 2025]. In this way, ORIGIN aligns with the core philosophy of the pre-training of CLIP, which explicitly models embedding relationships of individual image-text pairs via contrastive learning, thus guiding the adaptation of VLMs effectively to enhance the its inference capability under distribution shifts.

While we only adapt the InfoNCE loss at image-to-text alignment level, our framework can be naturally extends to the full InfoNCE loss adopted for the pre-training of CLIP by adding the text-to-image alignment counterpart, e.g., under text-level distribution shifts.

3.3 Bridging Inference and Adaptation via OT

In this section, we further bridge the inference and adaptation objectives of VLMs theoretically through the lens of OT. Specifically, as shown in the following lemma, the soft-label InfoNCE loss in Eq. (7) is mathematically equivalent to the entropic Wasserstein OT formulation in Eq. (5).

Lemma 3.1. *Given* $\mathbf{C}_{i,j} = 1 - \mathbf{v}_i^\top \mathbf{t}_j$*, the term* $\mathrm{KL}(\mathbf{S} \| \tilde{\mathbf{P}}^{\theta})$ *in Eq.* (7) *can be formulated as follows*

$$\mathrm{KL}(\mathbf{S} \| \tilde{\mathbf{P}}^{\theta}) = \underbrace{\frac{1}{\tau}\left(\langle \mathbf{C}, \mathbf{S} \rangle - \tau H(\mathbf{S})\right)}_{\textit{entropic Wasserstein OT}} + \underbrace{1 + \frac{1}{b}\sum_{i=1}^{b} \log\left[b \sum_{k=1}^{c} \exp\left(\frac{\mathbf{v}_i^\top \mathbf{t}_k - 1}{\tau}\right)\right]}_{\textit{constant w.r.t. } \mathbf{S}} \tag{10}$$

Detailed proof of Lemma 3.1 can be found in Appendix A.1. We can see from Eq. (10) that $\mathrm{KL}(\mathbf{S} \| \tilde{\mathbf{P}}^{\theta})$ can be rewritten into a entropic Wasserstein OT term plus a constant w.r.t. the transport plan $\mathbf{S}$ which does not affect the optimization of $\mathbf{S}$. Therefore, optimizing Eq. (10) w.r.t. the transport plan $\mathbf{S}$ is equivalent to optimizing Eq. (5), i.e.,

$$\min_{\mathbf{S} \in \Pi(\boldsymbol{\mu}_i, \boldsymbol{\mu}_t)} \mathrm{KL}(\mathbf{S} \| \tilde{\mathbf{P}}^{\theta}) := \min_{\mathbf{S} \in \Pi(\boldsymbol{\mu}_i, \boldsymbol{\mu}_t)} \langle \mathbf{C}, \mathbf{S} \rangle - \tau H(\mathbf{S}) \tag{11}$$

with the scalar temperature τ in Eq. (7) becoming the entropic regularization weight ϵ in Eq. (5), i.e., $\epsilon := \tau$. In this way, we theoretically prove that the inference and adaptation of VLMs by the proposed ORIGIN share the same objective functions w.r.t. different optimized variables $\mathbf{S}$ and the parameter θ of the image encoder, implying that these two processes are mutually beneficial instead of decoupled from each other. For one thing (*inference for adaptation*), OT estimate robust soft pseudo-labels at the sample-level which guides the adaptation of VLMs effectively; for another (*adaptation for inference*), the InfoNCE loss aligns with the pre-training philosophy of CLIP, which explicitly models sample-level relationships across modalities and effectively guides the re-alignment of image and text embeddings under distribution shift, generating accurate predictions for improved inference. By alternate between optimizations of $\mathbf{S}$ and θ, ORIGIN achieves mutual benefits of OT-based inference and InfoNCE-based adaptation, leading to effective TTA of VLMs. Formally, the unified objective of inference and adaptation in ORIGIN at test time is introduced as follows.

$$\min_{\theta, \mathbf{S} \in \Pi(\boldsymbol{\mu}_i, \boldsymbol{\mu}_t)} \mathrm{KL}(\mathbf{S} \| \tilde{\mathbf{P}}^{\theta}) \quad \tilde{\mathbf{P}}^{\theta}_{i,j} := \frac{\exp(\mathbf{v}_i^{\top} \mathbf{t}_j / \tau)}{b \sum_{k=1}^{c} \exp(\mathbf{v}_i^{\top} \mathbf{t}_k / \tau)} \tag{12}$$

Note that the soft-label InfoNCE loss on the text side can also be unified with a entropic Wasserstein OT formulation in a similar fashion, making our unified theory of inference and adaptation of VLMs applicable to both adaptation settings under image and text-level distribution shifts.

4 Experiments

In this section, we carry out comprehensive experiments and analyses to evaluate the proposed ORIGIN from the following aspects:

- **Q1:** How effective is ORIGIN compared to state-of-the-art methods under distribution shifts?
- **Q2:** How does ORIGIN trade off effectiveness and efficiency?
- **Q3:** How does inference and adaptation of ORIGIN benefit each other under a unified OT view?
- **Q4:** How sensitive is ORIGIN to different hyperparameter and design choices?

Due to limited space, we include detailed experimental setup in Appendix B, including introductions of datasets & metrics, VLM backbones, TTA baselines, reproducibility details, and machines used for conducting all experiments.

4.1 Benchmarking Results

We first benchmark ORIGIN with ViT-B-16 visual encoder on three standard VLM TTA benchmarks, whose results are reported in Table 1. We observe that **(1) The proposed ORIGIN consistently achieve state-of-the-art performance across all three benchmarks.** Specifically, on datasets with low-resolution images, ORIGIN outperforms the best competitor by a large margin, with a 7.4% and 4.8% improvement in mean accuracy on CIFAR10-C and CIFAR100-C, respectively. For the larger-scale ImageNet-C dataset, ORIGIN also achieve the best mean accuracy with at least 2.4% performance improvement, demonstrating the effectiveness of ORIGIN. **(2) OT inherently produces accurate predictions without adapting VLM parameters.** Comparing the performance of OT inference *without* adaptation in Table 2, with that of training-free baselines (TDA, DMN, VTE, ZERO, ECALP, and TENT), shows that OT inference alone outperforms all training-free heuristics in mean accuracy on CIFAR10-C, revealing that OT itself can infer accurate predictions, thus generating robust pseudo-labels that leads to effective adaptation. **(3) The InfoNCE loss re-aligns image and text embeddings of the same class effectively under distribution shifts.** As shown in Figure 3, the proposed ORIGIN produces image embeddings that not only align closely with the text embeddings of the same class, but also show clear clustering structures in the embedding space. In contrast, coarse-grained surrogate loss functions of other baselines leads to entangled image embeddings which deviated from their corresponding text embeddings, harming the adaptation performance.

4.2 Efficiency Results

We study the effectiveness-efficiency trade-off of the proposed ORIGIN and other training-based TTA methods, including RoTTA, TPT, MEMO, WATT, MINT and BATCLIP, on CIFAR10-C and CIFAR100-C. The results on mean accuracy vs. average runtime per corruption, defined as the averaged total amount of time required for inference and adaptation over different corruption types, are shown in Figure 4. We can see that **ORIGIN achieve state-of-the-art effectiveness and**

Table 1: Benchmarking results with ViT-B-16. The **1st**/2nd best results are highlighted in **bold** and underline, respectively. Additional benchmarking results with ViT-B-32 are shown in Table 4

Dataset	Method	Noise			Blur				Weather				Digital				Mean
		Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elast.	Pixel.	JPEG	
CIFAR10-C	Source	37.7	42.0	54.2	71.8	40.9	68.1	74.0	74.1	77.1	70.6	84.0	62.7	53.6	47.2	59.0	61.2
	TDA	40.9	44.0	50.5	72.3	44.6	71.7	75.6	75.5	78.2	71.6	85.7	62.4	56.3	48.3	58.1	61.9
	DMN	43.0	47.4	53.8	73.5	40.0	71.9	76.3	74.5	76.2	71.0	84.2	60.2	54.4	47.2	57.0	62.4
	VTE	42.8	46.7	63.9	71.0	45.6	69.0	73.6	76.7	78.5	71.0	85.7	57.5	59.5	60.8	61.7	64.3
	ZERO	38.7	43.9	57.3	71.7	40.9	69.0	74.5	74.5	77.6	72.7	84.1	60.3	55.5	48.6	61.8	61.6
	ECALP	46.5	50.8	61.1	71.3	42.0	70.8	74.9	75.9	79.0	71.1	86.3	60.8	57.9	49.5	61.4	64.2
	TENT	15.2	18.0	38.2	81.7	21.3	76.3	82.2	84.0	81.8	80.4	90.2	80.2	63.4	58.5	56.5	61.4
	RoTTA	39.2	43.1	55.4	71.8	41.3	68.3	74.0	74.3	77.8	70.9	85.0	63.4	54.2	49.1	60.3	62.3
	TPT	38.1	42.4	60.7	73.3	44.9	69.5	75.8	76.1	78.5	71.9	85.6	62.4	58.9	55.2	62.5	64.0
	MEMO	38.1	41.5	55.4	72.1	41.1	68.9	73.9	75.1	77.7	71.7	84.5	62.0	56.0	49.0	61.3	62.3
	WATT	46.3	53.3	59.9	75.2	38.3	71.9	76.2	77.3	80.7	76.1	87.9	75.7	55.1	62.4	63.2	66.7
	MINT	54.3	58.5	65.0	77.0	49.0	77.9	79.0	82.0	81.3	77.2	89.5	74.3	61.7	61.4	64.3	70.0
	BATCLIP	62.3	65.0	66.6	79.8	55.9	80.4	81.6	82.0	83.9	80.4	88.5	81.5	69.3	63.1	67.7	73.5
	ORIGIN	**71.8**	**74.3**	**77.7**	**84.1**	**70.5**	**84.1**	**85.4**	**86.6**	**86.5**	**87.1**	**90.9**	**89.3**	**76.9**	**81.1**	**74.7**	**80.9**
CIFAR100-C	Source	19.8	21.3	24.9	42.7	20.3	43.5	48.3	48.4	49.7	41.4	56.7	34.5	29.1	23.8	32.4	35.3
	TDA	22.4	25.5	29.3	43.1	19.4	43.5	49.1	48.1	50.7	41.5	58.4	35.2	29.0	24.4	32.9	36.5
	DMN	21.9	25.4	25.1	41.1	16.3	43.4	49.1	45.6	48.4	39.8	57.9	32.4	25.9	23.4	30.8	35.1
	VTE	18.0	18.9	27.9	39.9	19.5	39.5	45.3	48.6	46.6	40.5	55.0	30.1	32.2	30.4	31.3	34.6
	ZERO	19.0	20.6	28.6	43.7	19.4	43.1	49.1	48.8	50.0	44.4	57.7	35.0	31.2	25.0	33.6	36.8
	ECALP	23.4	25.3	30.5	43.0	19.5	42.8	49.6	47.8	50.3	42.7	58.0	35.1	30.0	25.4	31.9	37.0
	TENT	7.4	7.7	8.8	**51.8**	8.3	**52.2**	53.8	52.5	36.7	48.0	63.1	52.9	36.5	39.5	38.2	36.7
	RoTTA	20.9	22.6	26.5	42.8	20.4	43.3	48.2	49.1	49.8	41.5	57.7	34.5	28.8	25.5	33.3	35.8
	TPT	18.3	19.5	27.6	43.6	19.6	42.3	48.2	48.8	49.1	42.0	57.6	33.1	31.2	27.3	32.9	36.3
	MEMO	19.8	22.4	27.6	43.6	20.0	44.4	49.1	50.5	50.0	43.4	58.9	34.6	30.4	25.4	33.7	36.7
	WATT	25.8	27.3	31.8	48.4	23.4	48.9	52.7	53.3	52.2	48.1	62.9	45.6	35.4	37.0	38.1	42.0
	MINT	27.9	30.4	35.8	48.9	26.1	47.3	53.3	52.6	52.5	48.4	**64.4**	43.9	35.2	32.6	35.2	42.2
	BATCLIP	25.5	28.4	34.2	50.0	26.6	48.7	**55.2**	51.9	51.5	48.3	62.3	45.9	34.4	33.0	36.8	42.5
	ORIGIN	**36.9**	**39.1**	**44.7**	50.6	**36.2**	51.4	54.4	**53.9**	**53.0**	**52.0**	61.7	**53.2**	**41.4**	**45.8**	**42.3**	**47.3**
ImageNet-C	Source	11.4	12.9	11.9	23.3	15.4	24.9	22.2	32.0	29.9	35.6	54.4	17.3	12.7	30.9	33.5	24.3
	TDA	12.3	14.4	14.8	24.2	16.8	26.4	23.7	33.1	32.5	38.5	55.2	19.3	14.5	33.9	34.8	26.7
	DMN	11.5	13.9	14.4	22.3	16.2	24.4	22.7	31.7	30.8	35.8	53.9	16.0	12.9	32.7	25.3	24.0
	VTE	9.2	10.8	11.0	24.4	14.6	24.8	25.3	35.1	32.1	37.9	55.1	15.9	14.7	**38.8**	33.9	25.4
	ZERO	10.3	11.4	11.3	24.7	14.6	24.3	22.8	32.7	30.4	36.7	54.5	17.2	13.4	34.4	32.5	24.4
	ECALP	13.3	15.2	13.6	22.1	15.1	25.6	23.4	30.8	30.2	35.6	51.9	18.5	13.1	33.6	33.6	25.4
	TENT	5.5	5.3	7.9	25.5	19.5	26.8	24.3	33.8	30.8	38.3	54.6	22.6	13.9	34.8	35.6	25.2
	RoTTA	10.9	12.7	12.5	23.6	15.8	24.9	22.4	32.1	29.6	36.2	53.9	17.2	12.6	31.1	33.1	25.1
	TPT	8.1	9.4	10.4	23.6	15.2	24.9	24.4	34.4	32.3	36.9	55.1	16.1	14.2	34.1	33.8	24.7
	MEMO	10.6	12.0	11.9	23.7	15.5	24.7	23.2	32.9	29.8	35.9	54.3	17.3	12.3	31.6	33.0	25.1
	WATT	11.4	13.1	13.3	25.6	18.4	26.8	25.0	33.4	29.8	37.9	53.9	21.4	15.4	33.2	34.8	25.9
	MINT	19.7	19.9	19.2	26.5	21.6	29.6	25.6	32.3	29.3	39.3	55.3	24.0	18.8	36.2	**38.2**	29.3
	BATCLIP	19.7	20.9	19.4	26.1	21.1	30.4	28.8	35.4	31.2	40.5	**55.6**	26.0	24.1	36.9	37.4	30.5
	ORIGIN	**23.0**	**24.3**	**23.7**	**27.6**	**26.6**	**31.3**	**31.4**	**38.0**	**34.1**	**40.6**	50.2	**32.3**	**31.5**	38.4	38.0	**32.9**

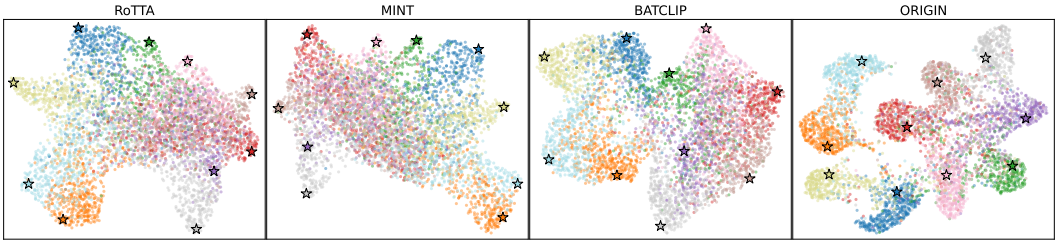

Figure 3: Embedding visualization of different VLM TTA methods by t-SNE on CIFAR10-C. Stars denote text embeddings of different classes, and round points denote image embeddings.

efficiency simultaneously, with the corresponding data points lying consistently on the upper-left region of both scatter plots. Specifically, ORIGIN outperforms the fastest baselines with up to 7.4% improvement in mean accuracy without additional latency. Compared with the remaining training-based methods, ORIGIN consistently outperforms all of them with up to 45× speed-up in runtime. This demonstrates that the additional overheads of solving the entropic Wasserstein OT during inference are practically negligible compared to the adaption of VLM parameters, which requires expensive forward and backward propagation, making ORIGIN surprisingly efficient.

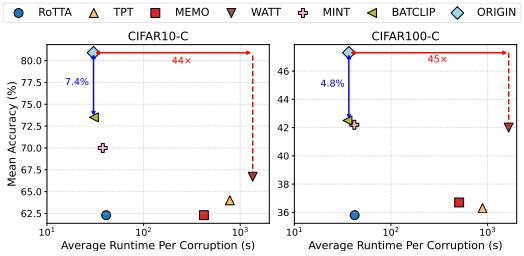

Figure 4: Accuracy vs. runtime trade-off of ORIGINand other training-based TTA methods, showing that the proposed ORIGIN achieves SOTA effectiveness and efficiency at the same time.

Table 2: Accuracy (%) of different variants of ORIGIN on CIFAR10-C.

Dataset	CIFAR10-C				
Corruption	Noise	Blur	Weather	Digital	Mean
EMB	44.69	63.54	76.47	55.83	60.13
OT	54.18	65.83	79.10	62.40	65.38
EMB+M	59.27	70.73	82.50	65.43	69.48
EMB+B	62.94	73.40	83.61	68.99	72.24
EMB+I	37.99	71.28	83.56	67.33	65.04
OT+M	62.31	71.89	82.95	67.82	71.24
OT+B	65.51	74.77	83.46	71.93	73.92
ORIGIN	74.43	81.05	87.60	80.44	80.88

4.3 Further Studies

4.3.1 Mutual Benefit of Inference and Adaptation

To isolate the individual effect of OT and InfoNCE loss at the inference and adaptation stage, and verify that they are mutually beneficial for VLM TTA, we compare the performance of ORIGIN against the following variants on CIFAR10-C:

- EMB / OT: Embedding-based / OT-based inference *without* adaptation.
- EMB+M: embedding-based inference with the adaption loss of MINT [Bao et al., 2025].
- EMB+B: embedding-based inference with the adaption loss of BATCLIP [Maharana et al., 2025].
- EMB+I: embedding-based inference with the InfoNCE loss.
- OT+I: OT-based inference with the adaption loss of MINT [Bao et al., 2025].
- OT+B: OT-based inference with the adaption loss of BATCLIP [Maharana et al., 2025].
- ORIGIN: OT-based inference with the InfoNCE loss.

The results are shown in Tables 2. **(1) OT-based inference predict reliable pseudo-labels which leads to consistent performance improvement.** Specifically, comparing the first three rows of Tables 2 with the last three rows reveals that OT-based inference consistently outperforms embedding-based inference under the same adaption loss, with an average performance improvement of 1.78%, 1.68%, and 15.84%. This demonstrates that OT consistently infer robust pseudo-labels during adaptation which leads to effective adaptation at test-time. **(2) InfoNCE loss improves the performance upperbound of TTA.** While class-level surrogate loss functions of MINT and BATCLIP sometimes avoid noise amplification inside pseudo-labels, the InfoNCE loss contributes to a higher performance upper-bound with an average performance improvement of 9.64% over MINT and 6.96% over BATCLIP when supervised by OT-induced pseudo-labels, demonstrating the necessity of explicit modeling of sample-level relationships across different modalities. **(3) OT-based inference and the InfoNCE-based adaptation are mutually beneficial under a unified formulation.** The proposed ORIGIN, which theortically bridges the inference and adaptation objectives, consistently achieves the best performance, demonstrating that OT-based inference and InfoNCE-based adaption inherently benefit each other.

4.3.2 Hyperparameter Analysis

We study the sensitivity of ORIGIN to three different hyperparameters on CIFAR10-C: the temperature τ in Eq. (7) (i.e., the entropic regularization weight ϵ in Eq. (5)), the EMA parameter α in Eq. (6), and the batch size b. **In general, the performance of ORIGIN remains stable under different choices of hyperparameters with consistent outperformance than its best competitors**. Detailed analysis for each parameter are provided as follows.

Temperature τ. We find that overly small or large τ typically leads to a slight performance degradation for ORIGIN. A small τ sharpens the softmax distribution and enforces deterministic OT prediction, but could potentially introduces errors due to overconfidence. A large τ avoids overly confident prediction, but leads to noisy supervision signals which limits the adaptation performance.

Temperature α. We can see that ORIGIN remain insensitive to different choices of α that trade-off historical and current text-level distribution, with less than 1% performance variation.

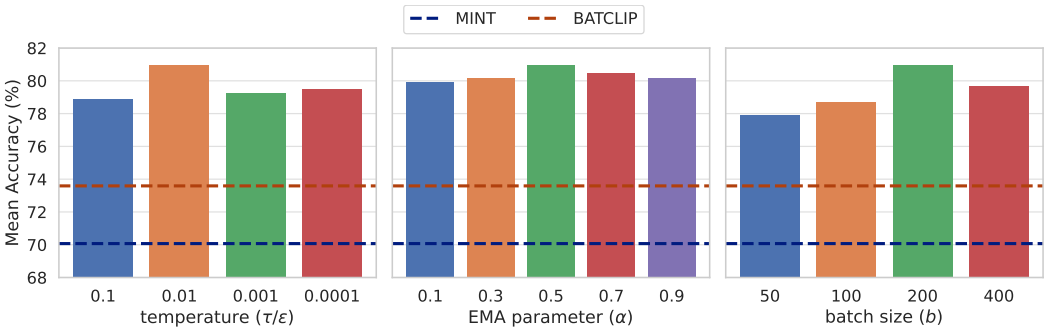

Figure 5: Hyperparameter sensitivity study of ORIGIN on CIFAR10-C

Batch size b. The influence of batch size b shares similar features as that of τ, where overly small or large b leads to performance degradation. Small b allows fine-grained adaption but provides limited global information for OT-based inference per batch. By contrast, large b leads to more accurate OT-based inference on a batch level but allows limited steps for adaption.

5 Related Works

VLM Test-time Adaptation. VLM TTA adapts pre-trained VLMs like CLIP [Maharana et al., 2025] from the source domain to target domain featured by distribution shifts at test time, without access to source training data or test labels. Early TTA methods for VLMs focus primarily on prompt optimization [Shu et al., 2022a, Xiao et al., 2025], memory management for historical samples [Karmanov et al., 2024, Zhang et al., 2024, Li et al., 2024b], or data augmentation & selection [Döbler et al., 2024], without adapting parameters inside different encoders of VLMs. While efficient, their performance are inherently limited without parameter adaptation. More recent efforts attempt to adopts normalization layers of the encoders inside VLMs by crafting coarse-grained surrogate loss functions. Specifically, BATCLIP [Maharana et al., 2025] adapts both the image and text encoders of CLIP by simultaneously separating image prototypes and aligning image-text prototypes. MINT [Bao et al., 2025] encourage inter-class separability based on averaged image embeddings. However, they overly relies on noisy-pseudo labels to provide coarse-grained supervision and fails to explicitly model cross-modal relationship at sample-level, leading to suboptimal adaptation performance.

Optimal Transport for VLMs. OT has attracted increasing attention in multimodal learning and vision-language models. PLOT [Chen et al., 2022] and Prompt-OT [Chen et al., 2026] leverages OT for prompt learning to improve VLMs on downstream vision-language tasks. SWAB [Yi et al., 2024] selects suitable VLMs for zero-shot classification by bridging the statistical gap across the open-source and target datasets via OT. OT-CLIP [Shi et al., 2024] leverages different variants of OT to improve the pretraining performance of CLIP, while OTCCLIP [Zhang et al., 2025] improves the robustness of CLIP against data poisoning with OT-based alignment at pretraining. While OT has been widely adopted for VLMs, none of existing work bridges the inference and adaptation of VLMs at test time through the lens of OT from a theoretical perspective.

6 Conclusion

In this paper, we study the VLM TTA problem by theoretically bridging the inference and adaptation of VLMs at test time through the lens of optimal transport to achieve their mutual benefits. To improve the robustness of pseudo-labels during inference, we reformulate the inference of VLMs into a Wasserstein OT formulation, generating reliable pseudo-labels at sample-level for effective adaptation. To leverage fine-grained supervision signals during adaptation, we adopt a soft-label InfoNCE loss to explicitly model cross-modal relationships at the sample-level via contrastive learning, effectively improving adaptation performance which further empowers robust inference at the same granularity. Finally, we theoretically unified the inference and adaptation objectives of VLMs, demonstrating that they inherently benefit each other rather than being decoupled stages. Extensive experiments demonstrate that ORIGIN significantly outperforms the best competitor in mean accuracy with state-of-the-art efficiency.

Acknowledgments and Disclosure of Funding

Use unnumbered first level headings for the acknowledgments. All acknowledgments go at the end of the paper before the list of references. Moreover, you are required to declare funding (financial activities supporting the submitted work) and competing interests (related financial activities outside the submitted work). More information about this disclosure can be found at: `https://neurips.cc/Conferences/2026/PaperInformation/FundingDisclosure`.

Do **not** include this section in the anonymized submission, only in the final paper. You can use the `ack` environment provided in the style file to automatically hide this section in the anonymized submission.

References

Wenxuan Bao, Ruxi Deng, and Jingrui He. Mint: A simple test-time adaptation of vision-language models against common corruptions. *arXiv preprint arXiv:2510.22127*, 2025.

Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Plot: Prompt learning with optimal transport for vision-language models. *arXiv preprint arXiv:2210.01253*, 2022.

Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. Graph optimal transport for cross-domain alignment. In *International Conference on Machine Learning*, pages 1542–1553. PMLR, 2020.

Xiwen Chen, Wenhui Zhu, Peijie Qiu, Hao Wang, Huayu Li, Haiyu Wu, Xuanzhao Dong, Aristeidis Sotiras, Yalin Wang, and Abolfazl Razi. Prompt-ot: An optimal transport regularization paradigm for knowledge preservation in vision-language model adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 667–676, 2026.

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.

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

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

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

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.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

Mingxuan Li et al. Test-time domain adaptation with positive and negative caches. In *CVPR*, 2024a.

Yushu Li, Yongyi Su, Adam Goodge, Kui Jia, and Xun Xu. Efficient and context-aware label propagation for zero-/few-shot training-free adaptation of vision-language model. *arXiv preprint arXiv:2412.18303*, 2024b.

Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096, 2022.

Sarthak Maharana, Baoming Zhang, Leonid Karlinsky, Rogerio Feris, and Yunhui Guo. Batclip: Bimodal online test-time adaptation for clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1569–1579, 2025.

Arkadi Nemirovski and Uriel Rothblum. On complexity of matrix scaling. *Linear Algebra and its Applications*, 302:435–460, 1999.

David Osowiechi, Mehrdad Noori, Gustavo A Hakim, Moslem Yazdanpanah, Ali Bahri, Milad Cheraghalikhani, Sahar Dastani, Farzad Beizaee, Ismail B Ayed, and Christian Desrosiers. Watt: Weight average test time adaptation of clip. *Advances in neural information processing systems*, 37:48015–48044, 2024.

Gabriel Peyré and Marco Cuturi. *Computational optimal transport: With applications to data science*. Now Foundations and Trends, 2019.

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.

Filippo Santambrogio. Optimal transport for applied mathematicians. 2015.

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

Liangliang Shi, Jack Fan, and Junchi Yan. Ot-clip: Understanding and generalizing clip via optimal transport. In *Forty-first International Conference on Machine Learning*, 2024.

Manli Shu, Weili Nie, De-An Huang, Zhiding Chen, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *NeurIPS*, 2022a.

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, 2022b.

Eloi Tanguy, Laetitia Chapel, and Julie Delon. Sliced optimal transport plans. *arXiv preprint arXiv:2508.01243*, 2025.

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

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. *arXiv preprint arXiv:2501.16404*, 2025.

Chao Yi, Yu-Hang He, De-Chuan Zhan, and Han-Jia Ye. Bridge the modality and capability gaps in vision-language model selection. *Advances in Neural Information Processing Systems*, 37: 34429–34452, 2024.

Qi Yu, Zhichen Zeng, Yuchen Yan, Zhining Liu, Baoyu Jing, Ruizhong Qiu, Ariful Azad, and Hanghang Tong. Planetalign: A comprehensive python library for benchmarking network alignment. *arXiv preprint arXiv:2505.21366*, 2025a.

Qi Yu, Zhichen Zeng, Yuchen Yan, Lei Ying, R Srikant, and Hanghang Tong. Joint optimal transport and embedding for network alignment. In *Proceedings of the ACM on Web Conference 2025*, pages 2064–2075, 2025b.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15922–15932, 2023.

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.

Tong Zhang, Kuofeng Gao, Jiawang Bai, Leo Yu Zhang, Xin Yin, Zonghui Wang, Shouling Ji, and Wenzhi Chen. Pre-training clip against data poisoning with optimal transport-based matching and alignment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9836–9849, 2025.

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.

Xingyu Zhu, Beier Zhu, Shuo Wang, Kesen Zhao, and Hanwang Zhang. Enhancing clip robustness via cross-modality alignment. *arXiv preprint arXiv:2510.24038*, 2025.

A Proof

A.1 Proof of Lemma 3.1

Lemma. *Given* $\mathbf{C}_{i,j} = 1 - \mathbf{v}_i^\top \mathbf{t}_j$*, the term* $\mathrm{KL}(\mathbf{S}\|\tilde{\mathbf{P}}^\theta)$ *in Eq.* (7) *can be formulated as follows*

$$\mathrm{KL}(\mathbf{S}\|\tilde{\mathbf{P}}^\theta) = \underbrace{\frac{1}{\tau}\left(\langle \mathbf{C}, \mathbf{S}\rangle - \tau H(\mathbf{S})\right)}_{\textit{entropic Wasserstein OT}} + \underbrace{1 + \frac{1}{b}\sum_{i=1}^{b} \log\left[b \sum_{k=1}^{c} \exp\left(\frac{\mathbf{v}_i^\top \mathbf{t}_k - 1}{\tau}\right)\right]}_{\textit{constant w.r.t. } \mathbf{S}}$$

Proof. Given that the transport plan $\mathbf{S}$ is a probability matrix with a uniform marginal distribution at the image side, i.e., $\sum_{i,j} \mathbf{S}_{i,j} = 1, \sum_{j=1}^{c} \mathbf{S}_{i,j} = \frac{1}{b}$, and $H(\mathbf{S}) := -\sum_{i,j} \mathbf{S}_{i,j} (\log \mathbf{S}_{i,j} - 1)$, the term $\mathrm{KL}(\mathbf{S}\|\tilde{\mathbf{P}}^\theta)$ in Eq. (7) can be expanded as follows

$$\begin{aligned}
\mathrm{KL}(\mathbf{S}\|\tilde{\mathbf{P}}^\theta) &= \sum_{i=1}^{b}\sum_{j=1}^{c} \mathbf{S}_{i,j} \log \frac{\mathbf{S}_{i,j}}{\tilde{\mathbf{P}}^\theta_{i,j}} \\
&= \sum_{i=1}^{b}\sum_{j=1}^{c} \mathbf{S}_{i,j} \log \mathbf{S}_{i,j} - \sum_{i=1}^{b}\sum_{j=1}^{c} \mathbf{S}_{ij} \left[\frac{\mathbf{v}_i^\top \mathbf{t}_j}{\tau} - \log b - \log \sum_{k=1}^{c} \exp\left(\frac{\mathbf{v}_i^\top \mathbf{t}_k}{\tau}\right)\right] \\
&= \sum_{i=1}^{b}\sum_{j=1}^{c} \mathbf{S}_{i,j} \log \mathbf{S}_{i,j} - \frac{1}{\tau}\sum_{i=1}^{b}\sum_{j=1}^{c} \mathbf{S}_{i,j} (1 - \mathbf{C}_{i,j}) + \log b + \sum_{i=1}^{b}\sum_{j=1}^{c} \mathbf{S}_{i,j} \log \sum_{k=1}^{c} \exp\left(\frac{\mathbf{v}_i^\top \mathbf{t}_k}{\tau}\right) \\
&= \frac{1}{\tau}\langle \mathbf{C}, \mathbf{S}\rangle - H(\mathbf{S}) + 1 - \frac{1}{\tau} + \log b + \sum_{i=1}^{b} \frac{1}{b} \log \sum_{k=1}^{c} \exp\left(\frac{\mathbf{v}_i^\top \mathbf{t}_k}{\tau}\right) \\
&= \frac{1}{\tau}(\langle \mathbf{C}, \mathbf{S}\rangle - \tau H(\mathbf{S})) + 1 + \frac{1}{b}\sum_{i=1}^{b} \log\left[b \sum_{k=1}^{c} \exp\left(\frac{\mathbf{v}_i^\top \mathbf{t}_k - 1}{\tau}\right)\right]
\end{aligned}$$

□

In this way, we have proven Lemma 3.1.

B Detailed Experimental Setup

Datasets & Metrics. We adopt three standard benchmarks for corrupted image classification: CIFAR10-C [Krizhevsky et al., 2009], CIFAR100-C [Hendrycks and Dietterich, 2019], and ImageNet-C [Deng et al., 2009], which share the same 15 different types of corruptions. Following the standard TTA protocol [Wang et al., 2020], all results reported are based on the highest severity level (Level 5). We adopt the classification accuracy as the evaluation metric [Maharana et al., 2025].

VLM Backbones. We adopt the CLIP [Radford et al., 2021] model with different visual encoders, including ViT-B-16 and ViT-B-32. ViT-B-16 is adopted as default unless noted otherwise.

VLM TTA Baselines. We compare ORIGIN against state-of-the-art VLM TTA approaches. For training-free methods which relies on memory management or data augmentation, we include TDA [Karmanov et al., 2024], DMN [Zhang et al., 2024], VTE [Döbler et al., 2024], ZERO [Farina et al., 2024], ECALP [Li et al., 2024b], and TENT [Wang et al., 2020]. For training-based methods that optimizes prompts or tunable parameters inside VLMs during adaptation, we include RoTTA [Yuan et al., 2023], TPT [Shu et al., 2022b], MEMO [Zhang et al., 2022], WATT [Osowiechi et al., 2024], MINT [Bao et al., 2025], and BATCLIP [Maharana et al., 2025].

Table 3: Detailed hyperparameter settings of ORIGIN.

Hyperparameters	α	τ/ϵ
CIFAR10-C	0.5	0.01
CIFAR100-C	0.5	0.01
ImageNet-C	0.3	0.01

Reproducibility. For all experiments, the reported results are averaged against 5 different runs with randomized data loader. To ensure a fair comparison, all baselines are evaluated on a dataset under the same batch size b, which is 200 for CIFAR10-C and CIFAR100-C, and 64 for ImageNet-C. Hyperparameters of all baselines VLM TTA methods are set as default in their official implementations. Detailed hyperparameter values of ORIGIN are listed in the following Table 3. For all training-based TTA methods, the AdamW optimizer is used with a learning rate of 5e-3 to adapt the VLMs.

Machine. All experiments are conducted on a server with dual Intel® Xeon® Gold 6240R CPUs and 4 NVIDIA Tesla V100-SXM2 32GB GPUs.

C Additional Experimental Results

Table 4: Benchmarking results with ViT-B-32. The **1st**/2nd best results are highlighted in **bold** and underline, respectively. The proposed ORIGIN consistently achieves the best performance across all benchmarks, with up to 10% improvement in mean accuracy.

Dataset	Method	Noise			Blur				Weather				Digital				Mean
		Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elast.	Pixel.	JPEG	
CIFAR10-C	Source	36.0	40.0	43.5	70.1	41.4	64.3	69.8	70.8	72.6	66.6	81.3	64.6	59.3	48.6	56.9	59.3
	TDA	41.8	43.2	41.7	71.3	44.7	66.9	72.4	72.6	74.7	68.4	83.5	66.1	62.9	50.9	55.8	61.1
	DMN	38.9	39.3	42.5	70.6	39.2	65.1	72.7	72.3	73.4	66.4	82.8	58.6	60.7	49.1	56.3	59.1
	VTE	47.7	49.8	53.6	71.6	54.2	67.9	72.9	76.9	76.5	70.9	83.7	60.9	69.2	58.7	61.2	65.3
	ZERO	36.9	41.6	46.8	70.8	42.1	65.3	71.7	71.9	73.4	68.8	83.0	65.3	62.4	47.7	59.0	60.4
	ECALP	44.4	47.3	44.2	70.8	43.2	67.9	74.2	72.7	74.7	67.7	82.2	62.9	61.5	47.7	55.8	60.9
	RoTTA	36.7	41.1	43.7	70.2	42.8	64.7	70.0	71.2	72.6	66.9	81.8	64.7	60.3	49.1	57.3	59.1
	TPT	43.6	46.8	48.0	71.1	47.8	66.5	72.0	73.5	76.4	68.6	84.0	66.4	64.1	51.4	58.1	62.4
	MEMO	36.7	40.6	44.1	71.2	42.3	64.7	70.2	72.1	73.4	68.1	82.2	64.4	61.0	47.5	57.4	59.6
	WATT	43.6	49.1	49.4	72.1	46.4	67.0	71.4	73.9	74.1	71.2	83.9	72.8	61.6	58.8	63.2	63.6
	MINT	54.4	57.8	47.5	73.3	55.9	74.0	76.1	74.5	74.6	70.5	85.6	70.9	64.5	59.2	59.4	66.4
	BATCLIP	51.8	56.0	54.2	76.0	55.2	74.6	75.5	76.8	78.5	75.2	86.3	77.0	67.1	57.8	61.6	68.2
	Ours	**68.0**	**72.1**	**65.4**	**81.3**	**70.2**	**81.1**	**83.4**	**83.1**	**83.5**	**84.2**	**88.4**	**86.3**	**74.5**	**80.3**	**73.3**	**78.2**
CIFAR100-C	Source	15.8	18.1	17.3	39.3	17.4	38.2	43.5	42.2	43.3	39.6	50.8	29.5	29.2	23.3	29.6	31.5
	TDA	18.7	20.9	19.8	40.2	17.9	39.7	45.1	44.3	45.2	40.3	52.2	30.9	29.6	23.0	31.0	33.7
	DMN	17.3	20.0	13.7	36.9	15.6	40.6	45.6	43.2	43.9	39.5	52.1	27.2	27.5	19.5	29.2	31.7
	VTE	16.7	18.1	19.1	40.0	22.9	38.7	43.7	45.0	45.2	39.0	49.8	28.6	34.4	27.1	30.2	33.4
	ZERO	15.5	18.2	19.7	41.3	17.2	40.0	44.3	43.4	43.4	40.8	51.4	29.3	30.4	21.5	31.5	32.2
	ECALP	18.7	19.2	17.4	38.3	17.8	39.9	44.3	43.7	44.9	41.0	52.7	31.2	30.7	23.1	30.1	33.2
	RoTTA	16.4	18.8	18.2	38.7	17.6	39.0	43.3	42.8	43.8	39.7	51.0	29.0	29.1	23.7	30.2	32.2
	TPT	16.5	17.5	17.0	39.1	19.3	38.8	43.9	43.4	44.8	40.3	50.7	27.6	30.5	23.8	29.3	32.4
	MEMO	16.9	18.1	18.1	41.3	17.2	40.6	45.5	44.3	44.4	40.5	52.6	29.7	30.9	22.6	31.3	33.1
	WATT	21.7	22.0	23.7	47.8	18.3	44.3	49.9	47.7	47.5	45.2	57.1	44.1	32.8	29.9	35.3	38.2
	MINT	23.7	26.3	21.5	46.0	23.5	42.3	48.5	46.5	45.2	43.3	55.7	38.6	33.8	28.2	32.1	37.2
	BATCLIP	21.5	25.0	22.6	46.6	22.9	44.4	49.5	47.0	46.5	44.7	58.1	38.6	34.6	28.6	33.4	37.3
	Ours	**32.7**	**35.4**	**30.7**	**49.8**	**34.9**	**48.6**	**53.2**	**50.2**	**50.9**	**50.0**	**58.8**	**49.6**	**41.2**	**41.5**	**39.8**	**44.2**
ImageNet-C	Source	13.2	13.1	12.4	24.5	11.4	23.1	20.6	25.3	26.3	30.5	50.9	17.0	19.1	32.1	29.5	23.2
	TDA	12.4	15.0	14.9	24.0	13.3	23.4	21.0	26.9	27.8	32.7	**51.8**	18.5	20.3	33.4	30.6	24.6
	DMN	13.5	14.1	14.7	23.3	12.0	21.8	19.4	23.4	26.1	29.3	50.7	14.8	20.3	32.1	29.4	23.3
	VTE	11.9	12.0	13.6	24.6	11.5	22.9	22.4	27.0	26.7	32.1	51.3	17.2	19.9	35.3	33.0	24.5
	ZERO	11.9	13.0	12.7	24.9	11.8	22.5	21.8	26.5	26.9	31.0	51.1	16.7	19.5	34.1	30.2	23.3
	ECALP	14.2	15.5	15.0	23.1	12.6	23.3	21.0	24.1	25.4	29.7	48.4	18.0	19.4	31.3	28.9	23.1
	RoTTA	13.5	13.7	13.5	24.2	12.5	23.0	20.0	25.5	26.4	30.2	50.0	17.6	19.4	32.1	29.5	23.5
	TPT	12.5	12.3	12.6	25.0	12.1	22.9	20.7	26.8	26.6	30.1	50.9	16.9	19.9	33.1	30.1	24.0
	MEMO	12.4	13.4	13.1	24.9	11.5	22.7	20.1	25.4	25.8	30.3	50.1	17.5	19.0	32.1	29.0	23.5
	WATT	14.5	14.2	14.2	26.1	15.5	25.3	22.3	26.4	25.4	31.2	50.3	22.6	19.4	33.1	30.3	24.7
	MINT	18.7	19.5	19.8	24.7	18.8	27.5	22.8	27.6	27.6	32.4	50.9	20.6	23.6	34.7	33.4	26.8
	BATCLIP	17.6	18.8	16.8	24.9	20.5	28.0	24.3	29.1	27.4	35.7	49.4	20.2	28.0	35.7	**34.5**	27.5
	Ours	**23.5**	**24.5**	**24.1**	**27.5**	**25.4**	**30.1**	**27.9**	**31.3**	**30.3**	**37.2**	45.7	**27.8**	**31.9**	**36.0**	34.0	**30.6**

D Limitations & Future Works

While ORIGIN demonstrates strong performance, we acknowledge a few limitations. Firstly, ORIGIN is an on line VLM TTA method which is not directly applicable to strictly episodic TTA settings without maintaining an explicit memory for historical samples. Secondly, like most training-based VLM TTA methods, ORIGIN requires access to model gradients for adapting VLMs at test

time, which may limits its applicability to closed-source proprietary models. For future works, it would be interesting to discover episodic extension of ORIGIN by leveraging efficient OT variants, e.g., sliced OT [Tanguy et al., 2025], for incremental updates of the transport plan, or extending the OT framework to black-box adaption settings.

E Boarder Impact

This paper aim to advance the field of cross-modal learning, vision-language model, and optimal transport. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the core contributions of our paper, which are fully supported by theoretical analysis and extensive experiments.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: Limitation and future works are included in Appendix D

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes].

   Justification: All theoretical results are included in Section 3.3 and proved in Appendix A, providing full set of assumptions and complete (and correct) proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Information needed to reproduce the experimental results are detailed in Appendix B

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code and data will be released upon publication.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Detailed training and test details are describe in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: the reported results are averaged against 5 different runs with randomized data loader.

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Computer resources are described in Appendix B

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform with the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Boarder impacts of our work are discussed in Appendix E

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our work focuses on VLM TTA and does not release new data or models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All creators and original owners of assets are properly credited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: Our paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.