Enhancing CLIP Robustness via Cross-Modality Alignment

Xingyu Zhu[1, 2] **Beier Zhu**[2] **Shuo Wang**[1†] **Kesen Zhao**[2] **Hanwang Zhang**[2]
[1]University of Science and Technology of China
[2]Nanyang Technological University
xingyuzhu@mail.ustc.edu.cn, shuowang.edu@gmail.com

Abstract

Vision-language models (VLMs) such as CLIP demonstrate strong generalization in zero-shot classification but remain highly vulnerable to adversarial perturbations. Existing methods primarily focus on adversarial fine-tuning or prompt optimization, they often overlook the gaps in CLIP's encoded features, which is shown as the text and image features lie far apart from each other. This misalignment is significantly amplified under adversarial perturbations, leading to severe degradation in classification performance. To address this problem, we propose **C**r**O**ss-moda**L**ity **A**lignment, dubbed **COLA**, an optimal transport-based framework that explicitly addresses adversarial misalignment by restoring both global image-text alignment and local structural consistency in the feature space. (1) COLA first projects adversarial image embeddings onto a subspace spanned by class text features, effectively filtering out non-semantic distortions while preserving discriminative information. (2) It then models images and texts as discrete distributions over multiple augmented views and refines their alignment via OT, with the subspace projection seamlessly integrated into the cost computation. This design ensures stable cross-modal alignment even under adversarial conditions. COLA is training-free and compatible with existing fine-tuned models. Extensive evaluations across 14 zero-shot classification benchmarks demonstrate the effectiveness of COLA, especially with an average improvement of 6.7% on ImageNet and its variants under PGD adversarial attacks, while maintaining high accuracy on clean samples.

1 Introduction

Vision-language models (VLMs) [17, 30] like CLIP [45] demonstrate strong generalization ability in zero-shot classification. However, they are highly susceptible to adversarial perturbations, where small but carefully crafted changes to input images can significantly mislead predictions [31, 32, 36]. Such vulnerabilities pose serious risks in critical applications such as medical diagnosis, autonomous driving, and security systems, where robustness and reliability are paramount.

Recent efforts to improve the adversarial robustness of VLMs can be broadly categorized into three directions: adversarial training [4, 62], which fine-tunes models with perturbed samples; prompt tuning [31, 63], which optimizes text input templates to resist attacks, and test-time defenses [21, 2, 59], which modify inputs or predictions on the fly. While these methods offer promising improvements, they suffer from high computational overhead or introduce substantial inference latency. More critically, they overlook a central issue: the misalignment between image and text modalities [70, 16]. This misalignment stems from CLIP's global matching paradigm, where the model is trained to align entire image embeddings with sentence-level textual embeddings. As shown in Figure 1(a), the text

†Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

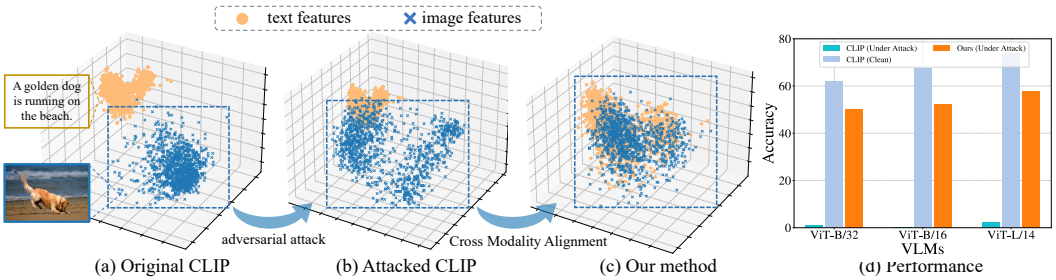


Figure 1: Visualization of image and text feature distributions under different conditions. We plot text and image embeddings via Principal component analysis (PCA) [1] and compare their performance. (a) Adversarial perturbations cause image features to scatter and misalign with text features. (b) Clean image and text features naturally form two distinct clusters as a result of contrastive training. (c) Our method mitigates the misalignment, making adversarial image features closer to the text features. (d) Classification performance across multiple VLMs shows that our method consistently improves robustness against adversarial inputs.

"*A golden dog running on the beach*" offers a fine-grained description that includes the object, its attributes, and the surrounding background. However, the image encoder processes the entire scene as a global representation, without explicitly modeling how these textual elements correspond to specific regions of the image. This results in image and text features being distributed independently in separate regions of the embedding space.

Such misalignment becomes particularly problematic under adversarial attacks [15, 31]. As illustrated in Figure 1(b), even small perturbations can distort the image embedding and severely disrupt global feature alignment, pushing visual representations away from their semantic prototypes. Beyond global shifts, attacks also damage the local structure within the feature space, causing nearby image embeddings to scatter and lose their internal consistency. As shown in Figure 1(d), this dual breakdown of alignment leads to a near-collapse in classification accuracy.

To address this issue, we propose a training-free framework that explicitly addresses adversarial misalignment at both the feature and semantic levels. First, we project adversarial image embeddings onto a text-induced subspace, eliminating non-semantic distortions and restoring feature space alignment. Then, we model images and texts as discrete distributions over multiple augmented views and refine their correspondence through optimal transport (OT) based on the projected features. Subspace projection is directly embedded into the OT cost, and we theoretically guarantee that it does not increase the transport distance. By jointly aligning feature embeddings and semantic distributions, our approach substantially improves the adversarial robustness of CLIP. Figure 1(d) illustrates the accuracy improvement over the attacked CLIP.

We conduct extensive experiments across 14 zero-shot classification benchmarks to evaluate the effectiveness of our method. Results demonstrate that COLA substantially improves adversarial robustness under multiple attack settings, with notable improvements such as an average gain of 6.7% under PGD and 4.8% under CW attacks on ImageNet and its variants, while maintaining high accuracy on clean images. Moreover, COLA can be directly applied to different CLIP fine-tuned models without any retraining, making it practical for real-world deployment.

2 Related Work

Adversarial robustness in VLMs. Adversarial robustness remains a fundamental challenge, as small, imperceptible perturbations can mislead model predictions [51, 7]. A common defense is adversarial training (AT) [33, 62, 47], which improves robustness but incurs high computational cost [50, 56]. Recent test-time defenses—such as generative purification [37, 61] and optimization-based methods [57, 35]—offer alternatives but often fail under adaptive attacks [12]. Hedge Defense (HD) [57], for example, perturbs inputs by maximizing cross-entropy loss, but requires an adversarially trained model. Meanwhile, several works explore CLIP's [45] robustness, noting its natural tendency to deflect attacks via counteractive perturbations in latent space. To further improve performance, researchers have applied adversarial fine-tuning [36, 54] and prompt tuning with frozen weights [31, 63].

These methods enhance robustness but rely on training. In contrast, our work proposes the first test-time defense for CLIP that is training-free, architecture-free, and efficient at inference. In contrast to prior efforts that rely on adversarial training, prompt tuning, or additional inference-time modules, we introduce a simple yet effective test-time defense for CLIP, which improves adversarial robustness by restoring image-text alignment through subspace projection and distribution-level matching, without requiring any model retraining or architectural changes.

Optimal transport. Optimal transport offers a principled way to compare probability distributions by capturing their geometric relationships [41]. With the development of efficient solvers such as the Sinkhorn algorithm [13], OT has been widely applied to tasks including generative modeling [3, 43, 44], domain adaptation [11], and structural alignment [9, 60]. In the vision-language domain, OT has enabled fine-grained alignment of image-text distributions in few-shot learning [27, 67], distribution calibration [22, 68], and prompt learning [8, 52]. Of particular relevance are recent OT-based methods for vision-language modeling [71], which improve zero-shot performance by enhancing alignment between visual and textual modalities. However, these approaches typically rely on training-time optimization or prompt tuning. While prior works focus on training-time alignment or require prompt tuning, our approach differs in that it introduces an efficient test-time OT framework for adversarially perturbed images. Specifically, we use OT to align projected image features with augmented textual prototypes, thereby enhancing robustness without any model fine-tuning.

3 Method

We propose a unified OT framework to enhance robust zero-shot classification by simultaneously addressing the modality misalignment caused by adversarial distortions and the mismatch between images and their text descriptions. We further provide theoretical guarantees that our method better preserves semantic similarity and yields larger margins, suggesting improved generalization.

3.1 Preliminaries

Zero-shot classification. CLIP [45] consists of a vision encoder $\Phi_v(\cdot)$ and a text encoder $\Phi_t(\cdot)$. Given a set of K class names $\{z_y\}_{y=1}^K$ and a hand-crafted template G, *e.g.*, ''`a photo of a` z_y'', the textual feature for class y is computed as $\mathbf{z}_y = \Phi_t(G(z_y))$. The visual feature for a testing samples x is calculated as $\mathbf{x} = \Phi_v(x)$, where both $\mathbf{x}$ and $\mathbf{z}_y$ lie in the same d-dimensional embedding space ($\mathbf{x}, \mathbf{z}_y \in \mathbb{R}^d$). CLIP performs classification by comparing the similarity between the visual feature $\mathbf{x}$ and all text prototypes $\{\mathbf{z}_y\}_{y=1}^K$:

$$y = \underset{y\in[K]}{\operatorname{argmax}}\, \mathbf{z}_y^\top \mathbf{x}. \tag{1}$$

Recent practices [42, 29, 48] replace the hand-crafted prompt $G(z_y)$ with a set of fine-grained class descriptions $\{\tilde{z}_y^m\}_{m=1}^M = \mathsf{LLM}(z_y)$ generated by large language models (LLMs). The corresponding text features $\{\mathbf{z}_y^m\}_{m=1}^M$ for class j are obtained via $\mathbf{z}_y^m = \Phi_t(\tilde{z}_j^m)$, and the average feature $\bar{\mathbf{z}}_y = \frac{1}{M}\sum_{m=1}^M \mathbf{z}_y^m$ is used in place of $\mathbf{z}_y$ in Eq. (1) for classification.

Adversarial perturbations. When the attacker has full access of model parameters, it becomes vulnerable to adversarial perturbations δ_a, which are typically generated via methods such as Projected Gradient Descent (PGD) [7]:

$$\delta_a = \arg\max_{\delta} L(x_i + \delta, y_i),\ \text{s.t. } \|\delta\|_p \le \epsilon_a \tag{2}$$

where y_i is the ground-truth and L is a loss function, typically cross-entropy. δ_a is constrained by an ℓ_p-norm budget ϵ_a, making it visually imperceptible yet highly effective at degrading accuracy.

3.2 Cross Modality Alignment under a Unified OT Framework

Original CLIP aligns clean images and texts into a unified feature space, but adversarial attacks on the visual modality severely disrupt this alignment. Moreover, visual features often capture background or irrelevant objects that are not reflected in LLM-generated descriptions, introducing

further semantic misalignment. In this work, we propose a unified OT framework to mitigate both types of misalignment by introducing feature space alignment and local semantics alignment.

Global feature alignment. Despite the contamination in image features, the subspace spanned by clean textual features serves as a reliable proxy for reconstructing the underlying clean image representations, a design inspired by [65]. Specifically, we arrange all class text embeddings $\{\mathbf{z}_y^m\}_{y,m}$ into a matrix $\mathbf{Z} \in \mathbb{R}^{d\times KM}$ and apply singular value decomposition (SVD) to extract the top-C principal components:

$$\mathbf{Z} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top, \quad \mathbf{U}_C = \mathbf{U}_{[:,1:C]}. \tag{3}$$

This defines a subspace $\mathcal{U} = \text{span}(\mathbf{U}_C)$, which captures C dominant directions shared across class embeddings. Since adversarial perturbations distort image features along directions away from $\mathcal{U}$, we project each perturbed image feature $\hat{\mathbf{x}}$ onto $\mathcal{U}$ to achieve alignment:

$$\Pi(\hat{\mathbf{x}}) = \mathbf{U}_C\mathbf{U}_C^\top\hat{\mathbf{x}}. \tag{4}$$

In Sec. 3.3, we show that the projection helps recover the pairwise similarity of clean image features.

Local structural alignment. While feature space alignment restores a unified space, projected image features can still misalign due to visual cues like background or irrelevant objects absent from LLM-generated text. To bridge this gap, we perform local semantics alignment for visual and textual representations. Specifically, for each adversarial image $\hat{x}$, we generate $N-1$ augmented views via random cropping, flipping, or resizing, and include the original to form a set $\{\hat{x}^n\}_{n=1}^N$, which are encoded into features $\{\hat{\mathbf{x}}^n\}_{n=1}^N$. Similarly, for each class name z_y, we obtain M textual descriptions by prompting LLMs to generate $M-1$ fine-grained variants in addition to the hand-crafted prompt, yielding features $\{\mathbf{z}_y^m\}_{m=1}^M$. We model each image and class as a discrete distribution, rather than a single embedding. For example, for an image $\hat{x}$ and class y, we model their distribution as:

$$\mathbb{P}(\mathbf{x}) = \sum_{n=1}^{N} a^n\delta(\hat{\mathbf{x}}^n - \mathbf{x}), \quad \mathbb{Q}_y(\mathbf{z}) = \sum_{m=1}^{M} b_y^m\delta(\mathbf{z}_y^m - \mathbf{z}), \tag{5}$$

where $\delta(\cdot)$ denotes the Dirac delta function, and a^n, b_y^m are the associated importance weights. To compute a^n for the augmented image feature $\hat{\mathbf{x}}^n$, we assess its entropy with respect to the average class embedding $\bar{\mathbf{z}}_y = \frac{1}{M}\sum_{m=1}^M \mathbf{z}_y^m$. Specifically, we define:

$$a^n = \frac{\exp\left(h(\hat{\mathbf{x}}^n)\right)}{\sum_{n'=1}^N \exp\left(h(\hat{\mathbf{x}}^{n'})\right)}, \quad h(\hat{\mathbf{x}}^n) = -\sum_{y=1}^{K} p(\bar{\mathbf{z}}_y|\hat{\mathbf{x}}^n)\log p(\bar{\mathbf{z}}_y|\hat{\mathbf{x}}^n). \tag{6}$$

The entropy $h(\hat{\mathbf{x}}^n)$ reflects the prediction confidence: views with lower entropy are assigned higher weights. The importance weights b_y^m for textual features are computed analogously.

Unified OT framework. Given the distributions $\mathbb{P}(\mathbf{x})$ and $\mathbb{Q}_y(\mathbf{z})$, the alignment between an adversarial image and each class is measured by the ot distance, which captures the minimal semantic matching cost between image and text features. We seek a transport plan $\mathbf{T}_y \in \mathbb{R}^{N\times M}$ that that moves mass from $\mathbb{P}(\mathbf{x})$ to $\mathbb{Q}_y(\mathbf{z})$, subject to the marginal constraints:

$$d_{\text{OT}}(\mathbb{P}(\mathbf{x}), \mathbb{Q}_y(\mathbf{z}); \mathbf{C}_j) = \min_{\mathbf{T}_y\geq\mathbf{0}} \langle \mathbf{T}_y, \mathbf{C}_y^\Pi\rangle, \quad \text{s.t.} \quad \mathbf{T}_y\mathbf{1}_M = \mathbf{a}, \quad \mathbf{T}_y^\top\mathbf{1}_N = \mathbf{b}_j, \tag{7}$$

where $\mathbf{a} = [a^1, \cdots, a^N]^\top$ and $\mathbf{b}_y = [b_y^1, \cdots, b_y^M]^\top$, and $\mathbf{1}_N$, $\mathbf{1}_M$ are all-ones vectors. $\mathbf{C}_y^\Pi \in \mathbb{R}^{N\times M}$ denotes the transportation cost between the N augmented image views and the M textual descriptions of class j, which is usually quantified using the cosine similarity. However, adversarial noise breaks alignment with text features, compromising the reliability of similarity measures. As a result, we design the OT cost matrix based our projected features. For image feature $\hat{\mathbf{x}}^n$, we compute:

$$\mathbf{C}_y^\Pi(n,m) = 1 - \cos\left(\Pi(\hat{\mathbf{x}}^n), \mathbf{z}_y^m\right), \tag{8}$$

where $\cos(\cdot,\cdot)$ denotes the cosine similarity. We classify by identifying the class y that yields the lowest transport cost:

$$y = \underset{y\in[K]}{\text{argmin}}\, d_{\text{OT}}(\mathbb{P}(\mathbf{x}), \mathbb{Q}_y(\mathbf{z}); \mathbf{C}_y^\Pi). \tag{9}$$

Section 3.3 demonstrates that our OT-based classifier achieves larger decision margins, indicating stronger generalization ability.

Table 1: Classification accuracy (%) on 9 widely-used datasets. The best and second best results are highlighted in **bold** and underline, respectively.

	Method	Pets		Flowers		Aircraft		DTD		Eurosat		Cars		Food		SUN		Caltech101	
		Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust
PGD Attacks	CLIP	87.4	1.0	65.5	1.1	20.1	0.0	40.6	3.0	42.6	0.0	52.0	0.0	83.9	0.7	58.5	1.1	85.7	14.7
	TeCoA	62.1	38.4	36.8	21.9	5.3	2.5	25.2	17.6	16.6	12.0	20.9	8.8	30.0	13.9	36.7	19.4	71.7	55.5
	PMG	65.9	41.2	37.0	23.4	5.6	2.2	21.8	15.0	18.5	12.6	25.4	11.7	36.6	18.6	38.0	22.6	75.5	61.1
	FARE	79.4	31.1	48.0	17.1	10.9	1.4	32.1	15.6	21.9	10.7	38.7	6.8	55.3	11.7	52.4	14.9	81.0	50.7
	RN	87.4	1.9	64.6	1.5	19.2	0.0	38.0	3.7	53.2	0.2	52.1	0.2	83.4	1.2	59.7	1.7	86.6	18.9
	TTE	**88.1**	50.3	65.2	35.9	20.2	6.2	**41.3**	23.9	44.4	6.9	52.7	22.4	84.0	43.9	59.1	30.8	85.8	67.6
	HD	80.9	12.0	58.2	7.3	16.4	1.3	34.9	11.6	39.1	4.6	44.3	2.7	80.3	8.0	53.2	6.4	82.3	31.5
	TTC	83.4	57.9	64.2	39.1	18.0	13.8	37.0	27.3	53.2	12.2	48.2	33.0	82.2	57.8	55.1	41.5	86.5	65.8
	COLA	87.9	**77.2**	**66.1**	**50.4**	**20.9**	**15.6**	41.0	**34.0**	**53.8**	**19.2**	**54.2**	**35.4**	**84.5**	**63.8**	**61.9**	**45.3**	**88.1**	**75.3**
CW Attacks	CLIP	87.4	1.6	65.5	1.4	20.1	0.0	40.6	2.9	42.6	0.0	52.0	2.4	83.9	1.1	58.5	1.8	85.7	20.9
	TeCoA	62.1	37.9	36.8	21.1	5.3	2.3	25.2	16.3	16.6	11.7	20.9	8.7	30.0	12.9	36.7	18.4	71.7	56.2
	PMG	65.9	39.3	37.0	21.3	5.6	1.9	21.8	13.7	18.5	11.9	25.4	10.5	36.6	16.6	38.0	20.4	75.5	61.6
	FARE	79.4	33.9	48.0	17.3	10.9	1.4	32.1	14.4	21.9	10.7	38.7	9.1	55.3	12.9	52.4	15.7	81.0	54.9
	RN	87.4	3.1	64.6	2.1	19.2	0.0	38.0	3.5	53.2	0.2	52.1	2.4	83.4	1.9	59.7	2.5	86.6	25.9
	TTE	**88.1**	51.1	65.2	35.0	20.2	5.2	**41.3**	22.6	44.4	6.4	52.7	21.2	84.0	44.6	59.1	29.4	85.8	69.4
	HD	80.6	13.8	57.8	8.5	16.2	1.0	34.9	10.1	40.1	3.5	43.6	5.1	81.0	9.8	54.1	7.9	83.0	36.3
	TTC	83.4	57.1	64.2	36.8	18.0	12.4	37.0	27.4	53.2	12.7	48.2	30.4	82.2	54.6	55.1	39.4	86.5	66.2
	COLA	87.9	**63.2**	**66.1**	**41.8**	**20.9**	**15.3**	41.0	**31.7**	**53.8**	**13.3**	**54.2**	**35.2**	**84.5**	**54.9**	**61.9**	**40.9**	**88.1**	**72.9**

3.3 Theoretical Analysis

Global feature alignment preserves pairwise similarity. We show that projection onto the subspace $\mathcal{U}$ preserves pairwise similarity among adversarial features. Let $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$ be the adversarial counterparts of two clean features $\mathbf{x}_1$ and $\mathbf{x}_2$:

$$\hat{\mathbf{x}}_1 = \mathbf{x}_1 + \boldsymbol{\delta}_1, \quad \hat{\mathbf{x}}_2 = \mathbf{x}_2 + \boldsymbol{\delta}_2, \tag{10}$$

where each perturbation $\boldsymbol{\delta}$ is decomposed as $\boldsymbol{\delta} = \boldsymbol{\delta}_{\parallel} + \boldsymbol{\delta}_{\perp}$, with $\boldsymbol{\delta}_{\parallel} \in \mathcal{U}$ and $\boldsymbol{\delta}_{\perp} \perp \mathcal{U}$. After projection, the adversarial feature satisfies:

$$\Pi(\hat{\mathbf{x}}) = \mathbf{x} + \boldsymbol{\delta}_{\parallel}. \tag{11}$$

Let Δ and Δ_{Π} denote the deviation from clean similarity before and after projection, respectively:

$$\Delta = |\cos(\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2) - \cos(\mathbf{x}_1, \mathbf{x}_2)|, \quad \Delta_{\Pi} = |\cos(\Pi(\hat{\mathbf{x}}_1), \Pi(\hat{\mathbf{x}}_2)) - \cos(\mathbf{x}_1, \mathbf{x}_2)|. \tag{12}$$

We show that projection yields lower cosine similarity distortion, *i.e.*, $\Delta_{\Pi} \leq \Delta$; the complete proof is given in Appendix A.1.

Our OT-based framework enjoys larger decision margins. Recall that the margin of an OT-based classifier with our projected cost matrix for a discrete distribution $\mathbb{P}(\mathbf{x})$ and its label y is defined as:

$$\gamma(\mathbf{C}^{\Pi}) = \min_{y' \in [K]} d_{\mathrm{OT}}(\mathbb{P}(\mathbf{x}), \mathbb{Q}(\mathbf{z}^{\Pi}_{y'}); \mathbf{C}_{y'}) - d_{\mathrm{OT}}\big(\mathbb{P}(\mathbf{x}), \mathbb{Q}(\mathbf{z}_y); \mathbf{C}^{\Pi}_{y}\big), \tag{13}$$

which measures the gap between the OT distance to the true class and the closest competing class. Let $\mathbf{C}_y(n, m) = 1 - \cos(\hat{\mathbf{x}}^n, \mathbf{z}^m_y)$ denote the cost matrix using the original perturbed features, and let $\gamma(\mathbf{C})$ be the corresponding OT classifier margin. We show that the margin of our OT classifier is larger than the original one: $\gamma(\mathbf{C}^{\Pi}) > \gamma(\mathbf{C})$. The detailed proofs are provided in Appendix A.2. Since classifiers with larger margins imply better generalization [5, 55], our approach leads to improved robustness against adversarial perturbations.

4 Experiments

In this section, we present the experimental results of our method under adversarial perturbations, including performance comparisons, ablation studies, and visualization analyses.

Table 2: Classification accuracy (%) on ImageNet and its variants datasets. The best results are highlighted in **bold**.

Method		ImageNet		ImageNet-A		ImageNet-V2		ImageNet-R		ImageNet-Sketch		AVG
		Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	
PGD	CLIP	62.1	1.1	30.3	0.0	55.0	0.8	65.3	6.1	39.7	5.01	26.5
	TTC	51.7	40.0	29.6	15.4	49.3	34.4	61.4	48.5	35.1	24.4	38.9
	COLA	**62.8**	**50.0**	**31.8**	**22.7**	**55.4**	**43.2**	**65.7**	**55.6**	**39.4**	**29.8**	**45.6**
CW	CLIP	62.1	1.1	30.3	0.1	55.0	1.1	65.3	6.8	39.7	5.4	26.6
	TTC	51.7	38.4	29.6	13.7	49.3	31.5	61.40	46.5	35.3	30.8	38.8
	COLA	**62.8**	**42.3**	**37.3**	**15.1**	**55.4**	**34.5**	**65.7**	**49.3**	**40.4**	**33.2**	**43.6**

Table 3: Classification accuracy (%) on 9 datasets under PGD attacks. The best results are highlighted in **bold**.

Method	Pets		Flowers		Aircraft		DTD		Eurosat		Cars		Food		SUN		Caltech101	
	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust
TeCoA	62.1	38.4	36.8	21.9	5.3	2.5	25.2	17.6	16.6	12.0	20.9	8.8	30.0	13.9	36.7	19.4	71.7	55.5
+ TTC	68.0	44.1	36.7	25.1	5.5	2.9	25.2	17.9	16.6	12.7	20.4	12.0	29.9	17.8	35.4	23.9	71.7	59.2
+ **COLA**	**69.2**	**54.9**	**37.0**	**31.3**	**6.4**	**4.7**	**26.5**	**18.4**	**16.9**	**14.0**	**32.2**	**28.3**	**30.8**	**22.0**	**38.9**	**27.3**	**73.5**	**61.1**
PMG	65.9	41.2	37.0	23.4	5.6	2.2	21.8	15.0	18.5	12.6	25.4	11.7	36.6	18.6	38.0	22.6	75.5	61.1
+ TTC	63.8	43.6	36.9	26.2	5.3	2.8	21.9	16.5	18.5	14.0	25.2	14.8	36.4	21.7	36.7	25.6	75.5	63.6
+ **COLA**	**66.0**	**46.2**	**37.8**	**29.0**	**5.8**	**4.5**	**24.5**	**19.0**	**19.7**	**14.9**	**26.9**	**16.8**	**37.9**	**25.1**	**40.8**	**29.1**	**77.5**	**66.6**
FARE	79.4	31.1	48.0	17.1	10.9	1.4	32.1	15.6	21.9	10.7	38.7	6.8	55.3	11.7	52.4	14.9	81.0	50.7
+ TTC	75.7	51.4	47.9	29.6	10.3	5.4	31.3	23.4	21.9	15.6	36.7	20.0	54.7	31.8	49.2	33.3	80.9	68.0
+ **COLA**	**80.3**	**57.6**	**48.6**	**35.9**	**11.4**	**7.7**	**33.4**	**26.9**	**22.8**	**16.1**	**41.1**	**28.7**	**55.6**	**35.2**	**54.2**	**45.3**	**83.3**	**73.2**

4.1 Setup

Datasets. We evaluate our method on 14 classification datasets spanning a broad range of domains, including generic objects (ImageNet [14], Caltech101 [20]), scenes (SUN397 [58]), textures (DTD [10]), satellite imagery (EuroSAT [23]), and various fine-grained categories such as pets, cars, flowers, food, and aircraft (Pets [39], Cars [26], Flowers [38], Food101 [6], Aircraft [34]). To further assess robustness under distribution shifts, we include five ImageNet variants: ImageNetV2 [46], ImageNet-Sketch [53], ImageNet-A [25], and ImageNet-R [24].

Implementation details. The attack budgets, including PDG attack and CW acctack [36, 7], are set of $\epsilon_a = 1/255$ in default. The number of steps for attacks is set as 10. All attacks are bounded by a L_∞ radius. For each test image, we generate $N = 5$ augmented views including the original. For each class, we use the LLM to generate $M = 50$ text descriptions. We select the top-$C = 256$ components from the SVD of class text features to build the projection matrix. All experiments are conducted on a single NVIDIA 3090 GPU if not specified.

Comparison methods. Our experiments are based on the pre-trained CLIP model, using ViT-B/32 as the visual encoder and a Transformer as the text encoder. We compare our method with test-time defences including Anti-Adversary (Anti-Adv) [2], Hedge Defence (HD) [57], Test-Time Transformation Ensembling (TTE) [40], and Test-Time Counterattacks (TTC) [59]. These methods are adapted to CLIP without additional networks. We also include fine-tuning-based baselines: TeCoA [36], PMG [54], and FARE [49], which adversarially fine-tune the vision encoder on TinyImageNet [28].

4.2 Main Results

Results on 14 datasets. We evaluate all methods assuming full access to model weights and gradients by the attacker. Table 1 reports classification accuracy on clean and adversarially perturbed images across 9 diverse datasets. Fine-tuning-based methods such as TeCoA [36], PMG [54], and FARE [49]

Table 4: Classification accuracy (%) on ImageNet and its variants datasets under PGD attacks. The best results are highlighted in **bold**.

Method	ImageNet		ImageNet-A		ImageNet-V2		ImageNet-R		ImageNet-Sketch		AVG
	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	
TeCoA	36.0	19.0	6.2	1.6	30.3	15.3	38.8	23.6	16.7	10.3	19.7
+ TTC	33.9	23.8	6.1	2.8	29.1	19.5	37.6	29.9	15.9	11.3	20.9
+ COLA	**36.6**	**27.4**	**6.7**	**3.7**	**31.3**	**23.0**	**39.0**	**32.6**	**16.9**	**14.6**	**23.1**
PMG	37.3	22.1	5.7	2.1	32.1	18.2	41.0	27.9	19.5	13.2	21.9
+ TTC	35.4	25.1	5.6	3.2	31.2	20.9	40.8	33.2	18.7	18.9	23.3
+ COLA	**37.8**	**30.2**	**6.3**	**4.3**	**33.1**	**25.2**	**42.4**	**38.3**	19.3	**19.9**	**25.6**
FARE	50.4	14.0	11.7	1.0	43.0	11.3	54.8	23.1	29.3	13.5	25.2
+ TTC	44.8	31.5	11.2	6.0	40.0	26.3	52.6	41.6	27.8	21.7	30.3
+ COLA	**50.9**	**37.5**	**11.9**	**7.5**	**44.5**	**29.6**	**55.3**	**45.2**	**29.6**	**25.5**	**33.7**

Table 5: Accuracy (%) on 9-datasets, ImageNet, and its five variant datasets, evaluated using ViT-B/16 and ViT-L/14 under PGD attacks. The best results are denoted in **bold**.

Model	ViT-B/16						ViT-L/14					
	9-datasets		ImageNet		IN-Variants		9-datasets		ImageNet		IN-Variants	
	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust
CLIP	63.6	0.8	67.6	0.2	57.3	1.5	70.4	3.5	73.9	2.5	69.6	4.8
TTC	61.3	13.9	67.1	20.1	53.8	17.9	69.9	16.2	68.8	21.9	68.3	19.3
COLA	**64.0**	**20.7**	**68.9**	**32.1**	**57.9**	**24.5**	**72.2**	**24.6**	**74.1**	**57.7**	**70.3**	**25.4**

improve robustness but significantly degrade clean performance. TTC [59], which introduces test-time counterattacks, enhances robustness further but requires stronger counterattack budgets and adds inference complexity.

In contrast, our method consistently improves robustness across all datasets and attack types (PGD and CW), while maintaining competitive clean accuracy. For example, on datasets like Food and Caltech101, our approach achieves over +5% absolute gains in robust accuracy compared to TTC, with only marginal clean performance drops. Table 2 shows results on ImageNet and its challenging variants. Our method consistently outperforms both CLIP and TTC, with especially large robustness gains on ImageNet-A and ImageNet-R—exceeding +7% under PGD attacks.

Results on finetuned CLIP. Our method aligns adversarially perturbed image features with their corresponding textual features and can be flexibly integrated into adversarially fine-tuned models as a plug-and-play module, without requiring architectural changes or additional training. Table 3 and Table 4 present results on ImageNet and its variants under PGD attacks.

TTC [59] improves robustness by generating test-time counterattacks using the fine-tuned model, but it often requires stronger counterattack budgets and introduces additional inference overhead. In contrast, our method consistently improves robust accuracy across all settings while preserving or minimally affecting clean performance. Specifically, on the 9-dataset benchmark, our method improves robust accuracy by +16.5% on TeCoA and +5.0% on PMG over their respective baselines, and by +10.8% and +2.6% over their TTC-augmented variants. On more challenging ImageNet variants, our approach achieves the highest robust accuracy in all cases. particularly excelling on ImageNet-R and ImageNet-Sketch, where robustness improvements of over +10% are observed.

Results on different backbones. To assess the generality of our method, we evaluate its performance across two CLIP backbones: ViT-B/16 and ViT-L/14. As shown in Table 5, our approach consistently achieves superior robustness against PGD attacks compared to both CLIP and TTC across all datasets. On ViT-B/16, our method improves robust accuracy over TTC by up to 12.0%, with consistent gains across 9-datasets, ImageNet, and its variants. Notably, on ViT-L/14, we observe a substantial 35.8% gain in robust accuracy on ImageNet, alongside improvements on the other test domains. These results highlight the adaptability of our method to different model capacities and architectures, and confirm its effectiveness in enhancing adversarial robustness without compromising clean accuracy.

Table 6: Classification accuracy (%) on 9-datasets under PGD attacks. The best and second best results are highlighted in **bold** and underline, respectively.

	Method	Pets		Flowers		Aircraft		DTD		Eurosat		Cars		Food		SUN		Caltech101	
		Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust
$\epsilon_a = 4/255$	CLIP	87.4	0.0	65.5	0.0	20.1	0.0	40.6	0.1	42.6	0.0	52.0	0.0	83.9	0.0	58.5	0.0	85.7	0.6
	TeCoA	53.9	3.7	27.8	3.8	3.5	0.1	20.1	5.2	17.5	10.7	15.2	0.4	21.9	1.4	28.2	2.3	64.4	21.0
	PMG	56.7	5.1	28.9	4.3	3.2	0.1	17.3	5.2	19.2	10.4	16.8	0.4	28.0	2.1	29.9	3.2	69.1	25.0
	FARE	70.1	0.3	41.0	0.6	7.8	0.0	28.0	2.5	18.2	7.3	32.1	0.0	42.0	0.2	43.6	0.6	76.6	10.1
	RN	87.4	0.0	64.6	0.0	19.2	0.0	38.0	0.1	53.2	0.0	52.1	0.0	83.4	0.0	59.7	0.0	86.6	0.7
	TTE	**88.1**	3.2	65.2	3.5	20.2	0.4	**41.4**	7.2	44.4	0.1	52.7	1.5	84.0	5.3	59.1	6.0	85.8	30.2
	HD	80.9	0.0	58.2	0.0	16.4	0.0	34.9	0.2	39.1	0.2	44.3	0.0	80.3	0.0	53.2	0.0	82.3	1.3
	TTC	64.7	24.6	63.2	13.6	16.0	6.4	35.7	11.4	53.2	13.6	41.5	12.8	80.0	17.9	46.7	13.4	86.2	36.7
	COLA	87.9	**29.2**	**66.1**	**29.3**	**20.9**	**6.9**	41.0	**22.0**	**53.8**	**23.3**	**54.2**	**18.4**	**84.3**	**24.3**	**61.9**	**24.2**	**88.1**	**43.8**

Table 7: Classification accuracy (%) of different models on 9-datasets, ImageNet, and its variant datasets under PGD and CW attacks. The best results are highlighted in **bold**.

Method	PGD Attacks						CW Attacks					
	9-datasets		ImageNet		IN-Variants		9-datasets		ImageNet		IN-Variants	
	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust
CLIP	59.5	2.4	62.1	1.1	47.6	3.0	59.5	3.5	62.1	1.1	47.6	3.3
OT w. $\mathbf{C}$	60.7	44.9	62.5	46.3	47.9	33.2	61.6	35.2	62.3	34.5	48.0	30.7
OT w. $\mathbf{C}^{\Pi}$	**62.0**	**46.2**	**62.8**	**50.0**	**48.0**	**37.8**	**62.8**	**42.3**	**62.8**	**42.3**	**49.7**	**33.0**

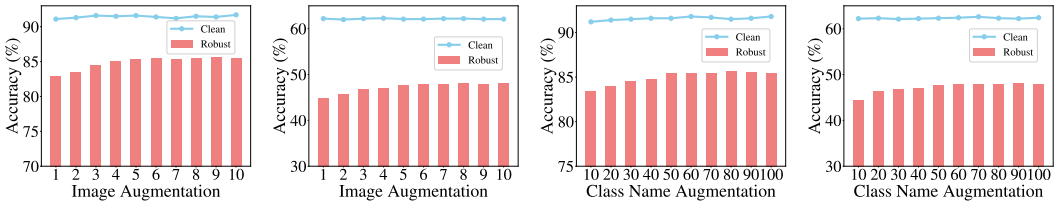
Results on large attack budgets. We further evaluate the robustness of all methods under a stronger adversarial budget of $\epsilon_a = 4/255$. As shown in Table 6, the performance of all baseline models drops sharply, with most robust accuracies approaching zero. This highlights their vulnerability under high-strength attacks. In contrast, our method maintains significantly higher robust accuracy across all nine datasets, demonstrating strong resistance to adversarial degradation. Notably, our approach achieves over 50% absolute gains in robust accuracy on datasets like Food, SUN, and Caltech101 compared to TTC [59], and outperforms all baselines by a large margin under this challenging setting.

4.3 Ablation Study

Effectiveness of the projected cost. Table 7 presents an ablation study comparing the standard CLIP model, the OT alignment with the original cost matrix $\mathbf{C}$, and our proposed projection-based cost matrix $\mathbf{C}^{\Pi}$. We observe consistent improvements when applying the subspace projection, indicating its effectiveness in mitigating adversarial perturbations. Specifically, across both PGD and CW attacks, $\mathbf{C}^{\Pi}$ achieves higher robust accuracy than both CLIP and the unprojected OT baseline. On the 9-datasets benchmark, the robust accuracy improves from 2.4% (CLIP) to 46.2%, while clean accuracy is also preserved. Similar trends are observed on ImageNet and its variants, with $\mathbf{C}^{\Pi}$ yielding up to over 3% gain in robust accuracy over $\mathbf{C}$ under PGD attakcs.

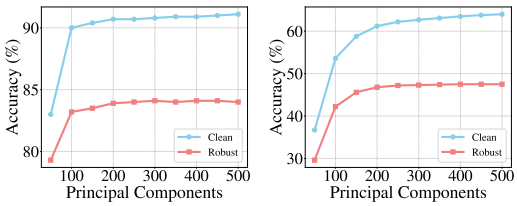
Effects of the number of augmentations. To investigate the sensitivity of our method to augmentation strategies, we evaluate the effect of varying the number of image and class name augmentations on both clean and robust accuracy, as shown in Figure 2. Increasing the number of image augmentations consistently enhances robustness, while the clean accuracy remains stable. However, the improvement becomes marginal when the number exceeds 5. A similar saturation effect is observed in class name augmentation, where performance gains plateau beyond 50 augmentations. These observations indicate that our method is robust to the choice of augmentation hyperparameters.

Effects of projection matrix construction. We study how varying the number of singular vectors C used to construct the projection matrix affects classification accuracy. As shown in Figure 3, increasing C steadily improves performance on both clean and adversarial examples across Caltech101 and ImageNet. The gains are more prominent when C is small and gradually saturate beyond $C = 200$,



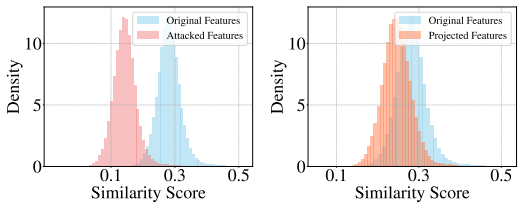
(a) Caltech101 (b) ImageNet (c) Caltech101 (d) ImageNet

Figure 2: Accuracy (%) comparisons across Caltech101 and ImageNet datasets with varying the number of augmentations. (a) and (b): Classification results under different numbers of image augmentations. (c) and (d): Classification results under different numbers of class name augmentations.



(a) Caltech101. (b) ImageNet.

Figure 3: Accuracy (%) with different numbers of principal components in the projection matrix.



(a) Attacked Similarity. (b) Projected Similarity.

Figure 4: Similarity distributions among original, attacked, and projected features on ImageNet.

especially for clean samples. In contrast, robust accuracy improves more slowly, suggesting that while additional components better preserve semantic structure for clean inputs, they provide limited benefit under adversarial perturbations. Based on this observation, we fix $C = 256$ in subsequent experiments to balance performance and efficiency.

To further understand the effectiveness of projection in Eq. (4), we analyze the similarity score distributions between image features and their corresponding text features under different conditions. As shown in Figure 4a, adversarial perturbations significantly reduce the similarity between image and text features, indicating disrupted alignment. In contrast, Figure 4b shows that projecting the attacked features onto the text-induced subspace effectively restores their similarity to the original level. This demonstrates that our projection effectively corrects adversarial misalignment and strengthens semantic consistency across modalities, thereby enhancing robustness in classification.

Running time. We compare the inference-time efficiency of our method on ImageNet using CLIP ViT-B/32 with a batch size of 128 on a single NVIDIA 3090 GPU. As shown in Table 8, our method completes evaluation in 28 minutes, significantly faster than TTC (40 minutes) while achieving both higher clean (62.8% vs. 51.7%) and robust (50.0% vs. 40.0%) accuracy. This efficiency stems from the training-free nature of our approach, which avoids the costly iterative optimization required by TTC.

Table 8: Comparison of running time on ImageNet with ViT-B/32.

Model	Running Time	Accuracy	
		Clean	Robust
CLIP	10min	62.1	1.1
TTC	40min	51.7	40.0
COLA	28min	62.8	50.0

5 Limitation and Conclusion

Limitation. While COLA substantially improves adversarial robustness, it still inherits potential biases from the pre-trained vision-language backbone [69, 64, 66]. In particular, the text-induced subspace may encode dataset-specific priors [19, 18], limiting generalization to unseen linguistic or visual domains. Moreover, stronger defenses could provoke more adaptive attacks, suggesting the need for future research on resilience under adaptive adversaries and fairness-aware robustness.

Conclusion. By enhancing robustness against adversarial manipulation, COLA contributes to safer multimodal systems, especially in high-stakes applications such as autonomous driving and medical imaging. We present COLA, a training-free and theoretically grounded framework that improves the adversarial robustness of CLIP by addressing modality misalignment. COLA leverages subspace projection to restore global alignment and employs optimal transport to refine local semantic consistency. By embedding projection into the OT cost computation, it maintains cross-modal

alignment without retraining or architectural modification. Theoretical analyses show that COLA reduces cosine distortion and enlarges decision margins, thereby improving generalization. Extensive experiments across 14 benchmarks confirm that COLA consistently enhances zero-shot classification robustness while preserving clean accuracy.

Acknowledgments and Disclosure of Funding

This research is supported by the National Natural Science Foundation of Anhui (Grant No. 2508085MF143) and the advanced computing resources provided by the Supercomputing Center of the University of Science and Technology of China (USTC). Additional support was provided by the National Research Foundation, Singapore, under the NRF Investigatorship Award (NRF-NRFI10-2024-0004).

References

[1] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2010.

[2] Motasem Alfarra, Juan C Pérez, Ali Thabet, Adel Bibi, Philip HS Torr, and Bernard Ghanem. Combating adversaries with anti-adversaries. In *AAAI*, 2022.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017.

[4] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. In *IJCAI*, 2021.

[5] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *NeurIPS*, 2017.

[6] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In *ECCV*, 2014.

[7] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, 2017.

[8] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Plot: Prompt learning with optimal transport for vision-language models. In *ICLR*, 2022.

[9] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *CVPR*, 2019.

[10] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

[11] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *TPAMI*, 2016.

[12] Francesco Croce, Sven Gowal, Thomas Brunner, Evan Shelhamer, Matthias Hein, and Taylan Cemgil. Evaluating the adversarial robustness of adaptive test-time defenses. In *ICML*, 2022.

[13] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2013.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[15] Junhao Dong, Piotr Koniusz, Xinghua Qu, and Yew-Soon Ong. Stabilizing modality gap & lowering gradient norms improve zero-shot adversarial robustness of vlms. In *KDD*, 2025.

[16] Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Improving cross-modal alignment in CLIP. In *ICLR*, 2025.

[17] Junfeng Fang, Zac Bi, Ruipeng Wang, Houcheng Jiang, Yuan Gao, Kun Wang, An Zhang, Jie Shi, Xiang Wang, and Tat-Seng Chua. Towards neuron attributions in multi-modal large language models. In *NeurIPS*, 2024.

[18] Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. Alphaedit: Null-space constrained knowledge editing for language models. In *ICLR*, 2025.

[19] Junfeng Fang, Yukai Wang, Ruipeng Wang, Zijun Yao, Kun Wang, An Zhang, Xiang Wang, and Tat-Seng Chua. Safemlrm: Demystifying safety in multi-modal large reasoning models. *CoRR*, abs/2504.08813, 2025.

[20] 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 *CVPR Workshops*, 2004.

[21] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. In *ICLR*, 2017.

[22] Dandan Guo, Long Tian, He Zhao, Mingyuan Zhou, and Hongyuan Zha. Adaptive distribution calibration for few-shot learning with hierarchical optimal transport. In *NeurIPS*, 2022.

[23] 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 J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 12(7):2217–2226, 2019.

[24] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021.

[25] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021.

[26] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshops*, 2013.

[27] Michalis Lazarou, Tania Stathaki, and Yannis Avrithis. Iterative label cleaning for transductive and semi-supervised few-shot learning. In *ICCV*, 2021.

[28] Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 2015.

[29] Jinhao Li, Haopeng Li, Sarah Erfani, Lei Feng, James Bailey, and Feng Liu. Visual-text cross alignment: Refining the similarity score in vision-language models, 2024.

[30] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

[31] Lin Li, Haoyan Guan, Jianing Qiu, and Michael Spratling. One prompt word is enough to boost adversarial robustness for pre-trained vision-language models. In *CVPR*, 2024.

[32] Xiao Li, Wei Zhang, Yining Liu, Zhanhao Hu, Bo Zhang, and Xiaolin Hu. Language-driven anchors for zero-shot adversarial robustness. In *CVPR*, 2024.

[33] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2018.

[34] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013.

[35] Chengzhi Mao, Mia Chiquier, Hao Wang, Junfeng Yang, and Carl Vondrick. Adversarial attacks are reversible with natural supervision. In *ICCV*, 2021.

[36] Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. In *ICLR*, 2023.

[37] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In *ICML*, 2022.

[38] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008.

[39] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012.

[40] Juan C. Pérez, Motasem Alfarra, Guillaume Jeanneret, Laura Rueda, Ali K. Thabet, Bernard Ghanem, and Pablo Arbeláez. Enhancing adversarial robustness via test-time transformation ensembling. In *ICCVW*, 2021.

[41] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, pages 355–607, 2019.

[42] Sarah M. Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *ICCV*, 2023.

[43] Junxiang Qiu, Lin Liu, Shuo Wang, Jinda Lu, Kezhou Chen, and Yanbin Hao. Accelerating diffusion transformer via gradient-optimized cache. *CoRR*, abs/2503.05156, 2025.

[44] Junxiang Qiu, Shuo Wang, Jinda Lu, Lin Liu, Houcheng Jiang, and Yanbin Hao. Accelerating diffusion transformer via error-optimized cache. *CoRR*, abs/2501.19243, 2025.

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[46] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019.

[47] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *ICML*, 2020.

[48] Karsten Roth, Jae Myung Kim, A Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. Waffling around for performance: Visual classification with random words and broad concepts. In *CVPR*, 2023.

[49] Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust CLIP: unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In *ICML*, 2024.

[50] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *NeurIPS*, 2019.

[51] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014.

[52] Dongsheng Wang, Miaoge Li, Xinyang Liu, MingSheng Xu, Bo Chen, and Hanwang Zhang. Tuning multi-mode token-level prompt alignment across modalities, 2023.

[53] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019.

[54] Sibo Wang, Jie Zhang, Zheng Yuan, and Shiguang Shan. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In *CVPR*, 2024.

[55] Colin Wei, Jason Lee, Qiang Liu, and Tengyu Ma. On the margin theory of feedforward neural networks. 2018.

[56] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training, 2020.

[57] Boxi Wu, Heng Pan, Li Shen, Jindong Gu, Shuai Zhao, Zhifeng Li, Deng Cai, Xiaofei He, and Wei Liu. Attacking adversarial attacks as a defense. *arXiv preprint arXiv:2106.04938*, 2021.

[58] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[59] Songlong Xing, Zhengyu Zhao, and Nicu Sebe. Clip is strong enough to fight back: Test-time counterattacks towards zero-shot adversarial robustness of clip. *arXiv preprint arXiv:2503.03613*, 2025.

[60] Hongteng Xu, Dixin Luo, and Lawrence Carin. Scalable gromov-wasserstein learning for graph partitioning and matching. *NeurIPS*, 2019.

[61] Jongmin Yoon, Sung Ju Hwang, and Juho Lee. Adversarial purification with score-based generative models. In *ICML*, 2021.

[62] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *ICML*, 2019.

[63] Jiaming Zhang, Xingjun Ma, Xin Wang, Lingyu Qiu, Jiaqi Wang, Yu-Gang Jiang, and Jitao Sang. Adversarial prompt tuning for vision-language models. In *ECCV*, 2024.

[64] Beier Zhu, Jiequan Cui, and Hanwang Zhang. Robust fine-tuning of zero-shot models via variance reduction. In *NeurIPS*, 2024.

[65] Beier Zhu, Jiequan Cui, Hanwang Zhang, and Chi Zhang. Project-probe-aggregate: Efficient fine-tuning for group robustness. In *CVPR*, 2025.

[66] Beier Zhu, Kaihua Tang, Qianru Sun, and Hanwang Zhang. Generalized logit adjustment: Calibrating fine-tuned models by removing label bias in foundation models. In *NeurIPS*, 2023.

[67] Xingyu Zhu, Shuo Wang, Jinda Lu, Yanbin Hao, Haifeng Liu, and Xiangnan He. Boosting few-shot learning via attentive feature regularization. In *AAAI*, pages 7793–7801, 2024.

[68] 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. *CoRR*, abs/2507.03657, 2025.

[69] Xingyu Zhu, Beier Zhu, Yi Tan, Shuo Wang, Yanbin Hao, and Hanwang Zhang. Enhancing zero-shot vision models by label-free prompt distribution learning and bias correcting. In *NeurIPS*, 2024.

[70] Xingyu Zhu, Beier Zhu, Yi Tan, Shuo Wang, Yanbin Hao, and Hanwang Zhang. Selective vision-language subspace projection for few-shot CLIP. In *ACM Multimedia*, 2024.

[71] Yuhan Zhu, Yuyang Ji, Zhiyu Zhao, Gangshan Wu, and Limin Wang. Awt: Transferring vision-language models via augmentation, weighting, and transportation, 2024.

Contents

1 Introduction **1**

2 Related Work **2**

3 Method **3**

3.1 Preliminaries 3

3.2 Cross Modality Alignment under a Unified OT Framework 3

3.3 Theoretical Analysis 5

4 Experiments **5**

4.1 Setup 6

4.2 Main Results 6

4.3 Ablation Study 8

5 Limitation and Conclusion **9**

A Proofs **15**

A.1 Proof of Cosine Similarity Distortion Bound 15

A.2 Proof of OT-Margin Amplification 16

B External Results **17**

C Algorithm **17**

A Proofs

A.1 Proof of Cosine Similarity Distortion Bound

Let $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$ be clean feature vectors with $\|\mathbf{x}_i\| = 1$. The adversarial features are denoted by

$$\hat{\mathbf{x}}_i = \mathbf{x}_i + \boldsymbol{\delta}_i, \quad \|\boldsymbol{\delta}_i\| \leq \epsilon,$$

where $\boldsymbol{\delta}_i = \boldsymbol{\delta}_\parallel^i + \boldsymbol{\delta}_\perp^i$, with $\boldsymbol{\delta}_\parallel^i \in \mathcal{U}$ and $\boldsymbol{\delta}_\perp^i \perp \mathcal{U}$. Let $\Pi(\hat{\mathbf{x}}_i) = \mathbf{x}_i + \boldsymbol{\delta}_\parallel^i$ denote the projection of the adversarial feature onto the subspace $\mathcal{U}$. The cosine similarity between adversarial features and projected features:

$$\cos(\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2) = \frac{(\mathbf{x}_1 + \boldsymbol{\delta}_1)^\top (\mathbf{x}_2 + \boldsymbol{\delta}_2)}{\|\hat{\mathbf{x}}_1\| \cdot \|\hat{\mathbf{x}}_2\|}, \quad \cos(\Pi(\hat{\mathbf{x}}_1), \Pi(\hat{\mathbf{x}}_2)) = \frac{(\mathbf{x}_1 + \boldsymbol{\delta}_\parallel^1)^\top (\mathbf{x}_2 + \boldsymbol{\delta}_\parallel^2)}{\|\Pi(\hat{\mathbf{x}}_1)\| \cdot \|\Pi(\hat{\mathbf{x}}_2)\|}.$$

We define the cosine distortion quantities as

$$\Delta = \left|\cos(\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2) - \cos(\mathbf{x}_1, \mathbf{x}_2)\right|, \quad \Delta_\Pi = \left|\cos(\Pi(\hat{\mathbf{x}}_1), \Pi(\hat{\mathbf{x}}_2)) - \cos(\mathbf{x}_1, \mathbf{x}_2)\right|.$$

We now compare Δ and Δ_Π. Using a second-order Taylor approximation for small perturbations $\|\boldsymbol{\delta}_i\| \ll 1$, we obtain:

$$\cos(\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2) \approx (\mathbf{x}_1^\top \mathbf{x}_2 + \mathbf{x}_1^\top \boldsymbol{\delta}_2 + \boldsymbol{\delta}_1^\top \mathbf{x}_2 + \boldsymbol{\delta}_1^\top \boldsymbol{\delta}_2)(1 - \mathbf{x}_1^\top \boldsymbol{\delta}_1 - \frac{1}{2}\|\boldsymbol{\delta}_1\|^2)(1 - \mathbf{x}_2^\top \boldsymbol{\delta}_2 - \frac{1}{2}\|\boldsymbol{\delta}_2\|^2).$$

$$\begin{aligned}\cos(\Pi(\hat{\mathbf{x}}_1), \Pi(\hat{\mathbf{x}}_2)) \approx &(\mathbf{x}_1^\top \mathbf{x}_2 + \mathbf{x}_1^\top \boldsymbol{\delta}_\parallel^2 + (\boldsymbol{\delta}_\parallel^1)^\top \mathbf{x}_2 + (\boldsymbol{\delta}_\parallel^1)^\top \boldsymbol{\delta}_\parallel^2) \\ &\times (1 - \mathbf{x}_1^\top \boldsymbol{\delta}_\parallel^1 - \frac{1}{2}\|\boldsymbol{\delta}_\parallel^1\|^2)(1 - \mathbf{x}_2^\top \boldsymbol{\delta}_\parallel^2 - \frac{1}{2}\|\boldsymbol{\delta}_\parallel^2\|^2).\end{aligned}$$

To simplify analysis, assume $\boldsymbol{\delta}_1 = \boldsymbol{\delta}_2 = \boldsymbol{\delta}$ and $\boldsymbol{\delta}_\parallel^1 = \boldsymbol{\delta}_\parallel^2 = \boldsymbol{\delta}_\parallel$. Since $\mathbf{x}_i \in \mathcal{U}$ and $\boldsymbol{\delta}_\perp \perp \mathcal{U}$, we have $\mathbf{x}_i^\top \boldsymbol{\delta}_\perp = 0$ and thus $\mathbf{x}_i^\top \boldsymbol{\delta} = \mathbf{x}_i^\top \boldsymbol{\delta}_\parallel$.

Under this setting, both distortions simplify to:

$$\Delta \approx 2\mathbf{x}_1^\top \boldsymbol{\delta}_\parallel - 2(\mathbf{x}_1^\top \mathbf{x}_2)(\mathbf{x}_1^\top \boldsymbol{\delta}_\parallel) + \mathcal{O}(\epsilon^2),$$

$$\Delta_\Pi \approx 2\mathbf{x}_1^\top \boldsymbol{\delta}_\parallel - 2(\mathbf{x}_1^\top \mathbf{x}_2)(\mathbf{x}_1^\top \boldsymbol{\delta}_\parallel) + \mathcal{O}(\epsilon^2).$$

Thus, the first-order terms in Δ and Δ_Π are identical. However, in the general case where $\|\boldsymbol{\delta}_\perp\| > 0$, the norm of the full feature $\hat{\mathbf{x}}_i$ is larger than that of its projection $\Pi(\hat{\mathbf{x}}_i)$, reducing the cosine similarity in the unprojected case.

Using Cauchy-Schwarz and bounding terms:

$$|\Delta_\Pi| \leq 2\|\boldsymbol{\delta}_\parallel\|(1 + |\mathbf{x}_1^\top \mathbf{x}_2|), \quad |\Delta| \geq \frac{2\|\boldsymbol{\delta}_\parallel\|(1 + |\mathbf{x}_1^\top \mathbf{x}_2|)}{\sqrt{1 + \frac{\|\boldsymbol{\delta}_\perp\|^2}{\|\boldsymbol{\delta}_\parallel\|^2}}}.$$

Therefore, when the perturbation contains a non-zero orthogonal component ($\|\boldsymbol{\delta}_\perp\| > 0$), we have

$$\frac{\Delta_\Pi}{\Delta} \leq \sqrt{\frac{\|\boldsymbol{\delta}_\parallel\|^2}{\|\boldsymbol{\delta}\|^2}} < 1,$$

which shows that the cosine similarity distortion is strictly reduced by projecting the adversarial features onto the subspace $\mathcal{U}$.

General case. When $\boldsymbol{\delta}_1 \neq \boldsymbol{\delta}_2$, let $s = \mathbf{x}_1^\top \mathbf{x}_2$ and $s_\parallel = \Pi(\hat{\mathbf{x}}_1)^\top \Pi(\hat{\mathbf{x}}_2)$. A first-order Taylor expansion yields:

$$\cos(\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2) \approx s + \mathbf{x}_1^\top \boldsymbol{\delta}_2 + \mathbf{x}_2^\top \boldsymbol{\delta}_1 - s(\mathbf{x}_1^\top \boldsymbol{\delta}_1 + \mathbf{x}_2^\top \boldsymbol{\delta}_2),$$

$$\cos(\Pi(\hat{\mathbf{x}}_1), \Pi(\hat{\mathbf{x}}_2)) \approx s_\parallel + \mathbf{x}_1^\top \boldsymbol{\delta}_{2\parallel} + \mathbf{x}_2^\top \boldsymbol{\delta}_{1\parallel} - s_\parallel(\mathbf{x}_1^\top \boldsymbol{\delta}_{1\parallel} + \mathbf{x}_2^\top \boldsymbol{\delta}_{2\parallel}).$$

Hence, the projected distortion satisfies:

$$\Delta_\Pi \leq |s - s_\parallel| + \epsilon \left(\frac{1}{\|\Pi \mathbf{x}_1\|} + \frac{1}{\|\Pi \mathbf{x}_2\|} \right),$$

where the first term reflects the semantic deviation between clean and projected subspaces, and the ϵ term accounts for asymmetric perturbations. Since projection removes orthogonal noise, the overall distortion ratio becomes:

$$\frac{\Delta_\Pi}{\Delta} \lesssim \frac{1}{1 + |s|} < 1,$$

showing that projection consistently suppresses cosine similarity distortion even when perturbations differ in direction or magnitude.

A.2 Proof of OT-Margin Amplification

We prove that the projected OT margin satisfies $\gamma(\mathbf{C}^\Pi) \geq \gamma(\mathbf{C})$, as stated in the main text.

Proof. Since the projection $\Pi(\hat{\mathbf{x}}^n) = \mathbf{x}^n + \boldsymbol{\delta}_\parallel^n$ removes the orthogonal perturbation $\boldsymbol{\delta}_\perp^n \perp \mathcal{U}$, and each text prototype $\mathbf{z}_y^m$ lies in the subspace $\mathcal{U}$, the dot product is preserved:

$$(\hat{\mathbf{x}}^n)^\top \mathbf{z}_y^m = \big(\Pi(\hat{\mathbf{x}}^n)\big)^\top \mathbf{z}_y^m.$$

Meanwhile, the norm of the adversarial feature satisfies:

$$\|\hat{\mathbf{x}}^n\| = \sqrt{\|\Pi(\hat{\mathbf{x}}^n)\|^2 + \|\boldsymbol{\delta}_\perp^n\|^2} \geq \|\Pi(\hat{\mathbf{x}}^n)\|,$$

with equality if and only if $\boldsymbol{\delta}_\perp^n = 0$. Therefore, the cosine similarities before and after projection are:

$$\cos(\hat{\mathbf{x}}^n, \mathbf{z}_y^m) = \frac{\big(\Pi(\hat{\mathbf{x}}^n)\big)^\top \mathbf{z}_y^m}{\|\hat{\mathbf{x}}^n\| \|\mathbf{z}_y^m\|}, \quad \cos(\Pi(\hat{\mathbf{x}}^n), \mathbf{z}_y^m) = \frac{\big(\Pi(\hat{\mathbf{x}}^n)\big)^\top \mathbf{z}_y^m}{\|\Pi(\hat{\mathbf{x}}^n)\| \|\mathbf{z}_y^m\|}.$$

Since $\|\hat{\mathbf{x}}^n\| \geq \|\Pi(\hat{\mathbf{x}}^n)\|$, we conclude:

$$\cos(\Pi(\hat{\mathbf{x}}^n), \mathbf{z}_y^m) = \cos(\hat{\mathbf{x}}^n, \mathbf{z}_y^m) \cdot \frac{\|\hat{\mathbf{x}}^n\|}{\|\Pi(\hat{\mathbf{x}}^n)\|} \geq \cos(\hat{\mathbf{x}}^n, \mathbf{z}_y^m).$$

Hence, the projected cost matrix entry is smaller:

$$\mathbf{C}_y^\Pi(n, m) = 1 - \cos(\Pi(\hat{\mathbf{x}}^n), \mathbf{z}_y^m) \leq 1 - \cos(\hat{\mathbf{x}}^n, \mathbf{z}_y^m) = \mathbf{C}_y(n, m).$$

Given $\mathbf{C}_y^\Pi(n, m) \leq \mathbf{C}_y(n, m)$, we now compare the OT distances. The OT distance is defined as:

$$d_{\text{OT}}(\mathbb{P}, \mathbb{Q}; \mathbf{C}_y) = \min_{\mathbf{T}_y \geq 0} \langle \mathbf{T}_y, \mathbf{C}_y \rangle, \quad \text{s.t. } \mathbf{T}_y \mathbf{1}_M = \mathbf{a}, \quad \mathbf{T}_y^\top \mathbf{1}_N = \mathbf{b}_y.$$

For any feasible transport plan $\mathbf{T}_y$, we have $\langle \mathbf{T}_y, \mathbf{C}_y^\Pi \rangle \leq \langle \mathbf{T}_y, \mathbf{C}_y \rangle$. Let $\mathbf{T}_y^* = \arg\min \langle \mathbf{T}_y, \mathbf{C}_y \rangle$. Then:

$$d_{\text{OT}}(\mathbb{P}, \mathbb{Q}; \mathbf{C}_y^\Pi) \leq \langle \mathbf{T}_y^*, \mathbf{C}_y^\Pi \rangle \leq \langle \mathbf{T}_y^*, \mathbf{C}_y \rangle = d_{\text{OT}}(\mathbb{P}, \mathbb{Q}; \mathbf{C}_y).$$

Define the OT distance reduction:

$$\Delta d_y = d_{\text{OT}}(\mathbb{P}, \mathbb{Q}; \mathbf{C}_y) - d_{\text{OT}}(\mathbb{P}, \mathbb{Q}; \mathbf{C}_y^\Pi) \geq 0.$$

Next, we compare the cost reductions for the true class $y^\star$ and a competing class y_{neg}. Note that:

$$\mathbf{C}_y(n, m) - \mathbf{C}_y^\Pi(n, m) = \cos(\Pi(\hat{\mathbf{x}}^n), \mathbf{z}_y^m) - \cos(\hat{\mathbf{x}}^n, \mathbf{z}_y^m) = \cos(\hat{\mathbf{x}}^n, \mathbf{z}_y^m) \cdot \left(\frac{\|\hat{\mathbf{x}}^n\|}{\|\Pi(\hat{\mathbf{x}}^n)\|} - 1 \right).$$

The multiplicative factor $\frac{\|\hat{\mathbf{x}}^n\|}{\|\Pi(\hat{\mathbf{x}}^n)\|} - 1 \geq 0$ is shared across all classes. Since $\cos(\hat{\mathbf{x}}^n, \mathbf{z}_{y^\star}^m) \geq \cos(\hat{\mathbf{x}}^n, \mathbf{z}_{y_{\text{neg}}}^m)$, it follows that:

$$\mathbf{C}_{y^\star}(n, m) - \mathbf{C}_{y^\star}^\Pi(n, m) \geq \mathbf{C}_{y_{\text{neg}}}(n, m) - \mathbf{C}_{y_{\text{neg}}}^\Pi(n, m).$$

Table 9: Performance comparison under AutoAttack.

Model	9-datasets (Clean)	9-datasets (Robust)	ImageNet (Clean)	ImageNet (Robust)
CLIP	63.6	0.5	67.6	0.2
TTC	61.3	8.3	67.1	16.2
COLA	**64.0**	**18.9**	**68.9**	**29.6**

Algorithm 1 Pipeline of **COLA**

Input: Adversarial visual feature $\mathbf{x}$ with its N augmentations $\{\hat{\mathbf{x}}^n\}_{n=1}^N$, and class textual embeddings $\{\mathbf{z}_y^m\}_{y,m}$ from LLM-generated descriptions.

Global feature projection. Construct the textual embedding matrix $\mathbf{Z} \in \mathbb{R}^{d\times KM}$ and perform SVD: $\mathbf{Z} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. Extract top-$C$ principal components $\mathbf{U}_C = \mathbf{U}_{[:,1:C]}$ and project each adversarial image feature onto the text-induced subspace: $\Pi(\hat{\mathbf{x}}^n) = \mathbf{U}_C\mathbf{U}_C^\top\hat{\mathbf{x}}^n$.
Local semantic distribution modeling. Represent the image and class as discrete distributions: $P(\mathbf{x}) = \sum_n a^n\delta(\hat{\mathbf{x}}^n - \mathbf{x})$, $Q_y(\mathbf{z}) = \sum_m b_y^m\delta(\mathbf{z}_y^m - \mathbf{z})$. The importance weights are normalized by prediction entropy: $a^n \propto \exp(h(\hat{\mathbf{x}}^n))$, $b_y^m \propto \exp(h(\mathbf{z}_y^m))$, where $h(\cdot)$ is defined in Eq. (6).
Optimal transport alignment. For each image–text pair $(\hat{\mathbf{x}}^n, \mathbf{z}_y^m)$, compute the projected transport cost: $C_y^\Pi(n,m) = 1 - \cos(\Pi(\hat{\mathbf{x}}^n), \mathbf{z}_y^m)$. Obtain the OT distance by solving:

$$d_{\mathrm{OT}}\big(P(\mathbf{x}), Q_y(\mathbf{z}); C_y^\Pi\big) = \min_{\mathbf{T}_y\geq 0}\langle \mathbf{T}_y, C_y^\Pi\rangle, \text{ s.t. } \mathbf{T}_y\mathbf{1}_M = a,\ \mathbf{T}_y^\top\mathbf{1}_N = b_y.$$

Classification. Predict the label with minimal OT distance:

$$\hat{y} = \arg\min_{y\in[K]} d_{\mathrm{OT}}\big(P(\mathbf{x}), Q_y(\mathbf{z}); C_y^\Pi\big).$$

Hence, for the respective optimal transport plans $\mathbf{T}_{y^\star}^*$ and $\mathbf{T}_{y_{\text{neg}}}^*$, we have:

$$\langle \mathbf{T}_{y^\star}^*, \mathbf{C}_{y^\star} - \mathbf{C}_{y^\star}^\Pi\rangle \geq \langle \mathbf{T}_{y_{\text{neg}}}^*, \mathbf{C}_{y_{\text{neg}}} - \mathbf{C}_{y_{\text{neg}}}^\Pi\rangle.$$

Therefore:

$$\Delta d_{y^\star} \geq \Delta d_{y_{\text{neg}}}.$$

Finally, we analyze the margin difference:

$$\begin{aligned}\gamma(\mathbf{C}^\Pi) - \gamma(\mathbf{C}) &= \big[d_{\mathrm{OT}}(\mathbb{P},\mathbb{Q};\mathbf{C}_{y_{\text{neg}}}^\Pi) - d_{\mathrm{OT}}(\mathbb{P},\mathbb{Q};\mathbf{C}_{y_{\text{neg}}})\big] - \big[d_{\mathrm{OT}}(\mathbb{P},\mathbb{Q};\mathbf{C}_{y^\star}^\Pi) - d_{\mathrm{OT}}(\mathbb{P},\mathbb{Q};\mathbf{C}_{y^\star})\big] \\ &= -\Delta d_{y_{\text{neg}}} + \Delta d_{y^\star} = \Delta d_{y^\star} - \Delta d_{y_{\text{neg}}} \geq 0.\end{aligned}$$

Thus, $\gamma(\mathbf{C}^\Pi) \geq \gamma(\mathbf{C})$, with equality if and only if $\boldsymbol{\delta}_\perp^n = 0$ for all n. □

B External Results

Analysis. As shown in Table 9, our method achieves the best robustness under *AutoAttack*, reaching 18.9% on 9-datasets and 29.6% on ImageNet, far surpassing CLIP (0.5% / 0.2%) and TTC (8.3% / 16.2%). These results show that the proposed subspace projection effectively filters non-semantic adversarial noise, while OT-based alignment restores image–text consistency.

C Algorithm

The overall procedure of COLA is summarized in Algorithm 1, which outlines the projection-based alignment and OT-based matching steps for adversarially robust inference.

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 [NA] .
- [NA] 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 NA).

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 be asked to also 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 "[NA] " 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: COLA performs test-time cross-modality alignment by projecting adversarial features and matching them via optimal transport. The method is theoretically justified and empirically validated across diverse CLIP models and robustness benchmarks.

   Guidelines:

   - The answer NA 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 NA 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: The limitations have been discussed in Section **??**.

Guidelines:

- The answer NA 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: We have provided the proof related to our method in Section A.

   Guidelines:

   - The answer NA 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: We have provided the details of our COLA framework in Section 3 and included the implementation details in Section 4.

   Guidelines:

- The answer NA 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: [Yes]

   Justification: We have uploaded the codes in supplemental material.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.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://nips.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, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have introduced the testing details in Section 4.

   Guidelines:

   - The answer NA 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: [No]

   Justification: Given the vision-language models, the process of our method is deterministic. Running multiple times will not introduce randomness.

   Guidelines:

   - The answer NA 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: We have provided the GPU in Section 4.

   Guidelines:

   - The answer NA 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: We have reviewed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA 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: We have discussed the societal impact on Section **??**.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our method aims to improve model robustness, which poses no such risks to the best of our knowledge.

    Guidelines:

- The answer NA 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: The datasets used in this paper are publicly available and cited appropriately.

    Guidelines:

    - The answer NA 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: [NA]

    Justification: We do not release new assets.

    Guidelines:

    - The answer NA 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: [NA]

Justification: We do not involve such experiments.

Guidelines:

- The answer NA 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: [NA]

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

    Guidelines:

    - The answer NA 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 rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

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

    Guidelines:

    - The answer NA 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 (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.