# GRADUAL ADVERSARIAL TRANSITION ALIGNMENT FOR ROBUST VISION-LANGUAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Vision-language models such as CLIP exhibit strong zero-shot generalization but remain highly vulnerable to adversarial perturbations. Most existing adversarial fine-tuning methods primarily constrain the clean/reference state and the final adversarial state, leaving the progressive semantic drift between them unsupervised. A natural remedy is to supervise the intermediate states of iterative attacks, but this requires an additional forward–backward pass for every state, and real trajectories carry attack-specific local fluctuations. We therefore investigate whether attack trajectories exhibit a compact geometric structure that can be efficiently approximated. Our analysis reveals that, despite substantial local fluctuations, the cumulative representation displacement follows a stable dominant direction across attacks and perturbation settings. This finding motivates an endpoint-induced surrogate path that approximates the dominant clean-to-adversarial transition using only the two endpoints. Inspired by gradual domain adaptation, we discretize this path into a sequence of ordered intermediate states and enforce KL-based prediction consistency between adjacent states. The resulting Gradual Adversarial Transition Alignment (GATA) replaces a coarse endpoint constraint with a sequence of locally smaller predictive transitions, providing efficient transition-level supervision without explicitly modeling the full attack trajectory. Extensive experiments across diverse attacks, perturbation strengths, downstream datasets, and vision-language backbones demonstrate improved adversarial robustness while preserving a favorable clean–robust trade-off.

## 1 INTRODUCTION

Vision-language models such as CLIP (Radford et al., 2021) achieve strong zero-shot generalization across diverse downstream tasks (Wang et al., 2024), yet remain vulnerable to adversarial perturbations that can lead to incorrect predictions (Schlarmann et al., 2024; Li et al., 2024a). From the representation perspective, adversarial attacks can therefore be viewed as inducing a semantic drift that progressively moves the visual representation away from the region consistent with its original semantics. Improving adversarial robustness consequently requires suppressing such semantic drift while preserving the pretrained semantic structure.

Adversarial fine-tuning is widely used to improve the robustness of pretrained VLMs (Mao et al., 2022; Schlarmann et al., 2024). Despite employing diverse feature-, prediction-, and attention-level objectives (Wang et al., 2024; Yu et al., 2024; Lu et al., 2026), existing methods largely follow an endpoint-based supervision that constrains the clean/reference state and the final adversarial state. However, iterative attacks induce a progressive evolution through intermediate representations rather than an instantaneous endpoint shift. As illustrated in Figure 1, endpoint supervision constrains the clean and final adversarial states but does not explicitly model the semantic transition between them. This motivates a transition-level question: *can robustness benefit from modeling how representations evolve throughout the attack process?*

A direct way to obtain transition-level supervision is to use the intermediate states generated by iterative attacks such as PGD (Madry et al., 2017) and CW (Carlini & Wagner, 2017). However, explicitly processing the full trajectory introduces substantial computational and memory overhead, while the real trajectory may also contain short-range fluctuations specific to the attack algorithm



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

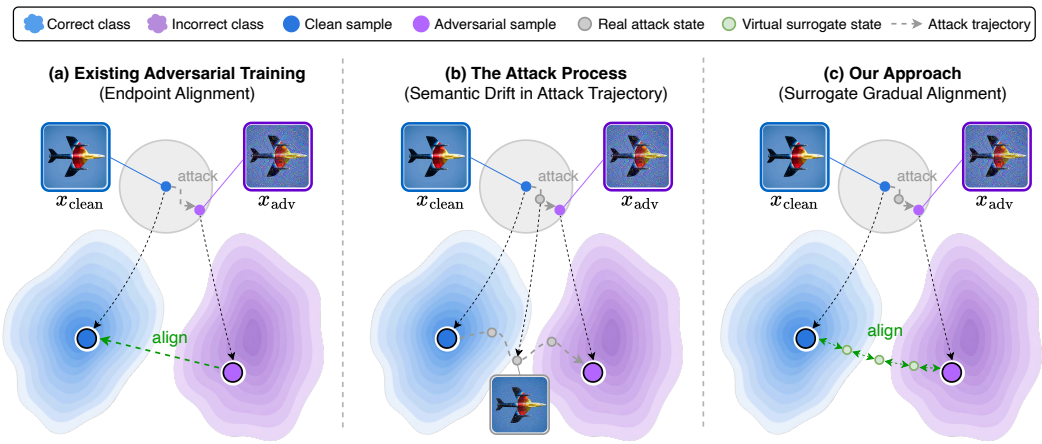



Figure 1: **Motivation of GATA.** (a) Existing adversarial fine-tuning applies endpoint-based supervision between the clean/reference and final adversarial states. (b) Real iterative attacks evolve through locally fluctuating intermediate states. (c) GATA approximates the principal transition with an endpoint-induced surrogate path and progressively aligns adjacent virtual states.

and optimization process. This raises a more fundamental question: *can the essential clean-to-adversarial transition be captured by a low-cost geometric surrogate?*

To answer this question, we analyze adversarial trajectories in CLIP's representation space from both local and global perspectives. Despite substantial directional fluctuations between consecutive updates, we observe that cumulative displacements remain considerably more stable and exhibit a persistent global direction across attack settings. This local–global contrast suggests that reproducing the full trajectory is unnecessary for capturing its principal transition. We therefore use the displacement between the clean and final adversarial representations as a compact geometric cue to construct an endpoint-induced surrogate path.

Given the surrogate path, the next question is how to exploit its transition structure for robust learning. Constraining only its endpoints would reduce again to conventional endpoint supervision. Inspired by gradual domain adaptation (GDA), which decomposes a large distribution shift into a sequence of smaller local transitions (Kumar et al., 2020; Wang et al., 2022; He et al., 2024), we propose Gradual Adversarial Transition Alignment (GATA). As illustrated in Figure 1(c), GATA replaces the coarse endpoint constraint with a sequence of local predictive transitions along the constructed surrogate path with little additional computational cost.

Our main contributions are summarized as follows:

- We revisit adversarial fine-tuning from a transition-level perspective, highlighting that endpoint-based supervision overlooks the progressive semantic drift between clean and adversarial states.
- We analyze the geometry of adversarial trajectories and uncover a stable global direction despite substantial local fluctuations, providing an empirical basis for approximating the principal adversarial transition from its endpoints.
- We propose GATA, which combines an endpoint-induced surrogate path with GDA-inspired gradual decomposition and local predictive alignment, enabling efficient transition-level supervision without explicitly processing real intermediate attack states.

## 2 Related Work

### 2.1 Adversarial Robustness of Vision-Language Models

Recent studies improve the adversarial robustness of vision-language models through prompt adaptation, training-free defenses, and adversarial fine-tuning (Li et al., 2024a; Sheng et al., 2025; Wang

et al., 2025; Xing et al., 2025; Tong et al., 2025; Zhang et al., 2025). Among these, adversarial fine-tuning adapts the pretrained visual encoder using adversarial examples while preserving zero-shot transferability. Existing methods employ diverse objectives, including feature-, attention-, and prediction-level alignment (Mao et al., 2022; Schlarmann et al., 2024; Wang et al., 2024; Li et al., 2024b; Yu et al., 2024; Dong et al., 2025a; Lu et al., 2026). Despite their different objectives and alignment spaces, most methods rely on clean-to-final-adversarial supervision. In contrast, we focus on the transition structure underlying attack-induced semantic drift and encourage the model to maintain semantic consistency throughout it.

### 2.2 Adversarial Trajectory and Path Modeling

Recent works have begun to exploit information beyond the final adversarial endpoint. Intermediate attack states have been used for robustness distillation (Dong et al., 2024), adversarial representation directions have been used for test-time defense (Liu et al., 2026), and adversarial simplices have been constructed from clean and intermediate attack states for robust alignment (Dong et al., 2025b). However, these methods primarily use trajectory-related information directly, without explicitly analyzing the geometric regularity of the attack trajectory itself. Moreover, relying on real intermediate attack states introduces additional computational cost. In contrast, we first characterize the local and global geometry of adversarial trajectories and use the resulting global regularity to construct an endpoint-induced surrogate transition, avoiding explicit supervision over the full real trajectory.

### 2.3 Gradual Domain Adaptation

Gradual domain adaptation (GDA) addresses large source-to-target distribution shifts by transferring knowledge through a sequence of intermediate domains (Kumar et al., 2020). Subsequent works show that both the number and arrangement of intermediate domains, as well as the transition path itself, can affect adaptation performance (Wang et al., 2022; He et al., 2024), while intermediate domains can also be recovered or explicitly constructed when unavailable (Chen & Chao, 2021; Zhuang et al., 2024). These works establish a general principle: a large distribution shift can be decomposed into a sequence of smaller local transitions. We adopt this principle for adversarial transition supervision, but differ from conventional GDA in that our intermediate states are explicitly constructed along a clean-to-adversarial representation path rather than observed target domains.

## 3 Preliminary

### 3.1 Zero-Shot Vision-Language Models

CLIP (Radford et al., 2021) maps images and textual class prompts into a shared embedding space. Let $f_\theta : \mathcal{X} \to \mathbb{R}^d$ and $g_\phi : \mathcal{T} \to \mathbb{R}^d$ denote the image and text encoders, respectively. For an image $x$ and a textual prompt associated with class $c$, their normalized representations are

$$v(x) = \frac{f_\theta(x)}{\|f_\theta(x)\|_2}, \qquad t_c = \frac{g_\phi(\mathrm{prompt}(c))}{\|g_\phi(\mathrm{prompt}(c))\|_2}, \tag{1}$$

where $v(x), t_c \in \mathbb{S}^{d-1}$ lie on the unit hypersphere. For a label set $\mathcal{C} = \{1, \dots, C\}$, the image-text similarity logit and zero-shot predictive distribution are

$$z_c(x) = \frac{v(x)^\top t_c}{\tau}, \quad p_\theta(y = c \mid x) = \frac{\exp(z_c(x))}{\sum_{j=1}^{C} \exp(z_j(x))}. \tag{2}$$

The predicted class is therefore $\hat{y}(x) = \arg\max_{c \in \mathcal{C}} p_\theta(y = c \mid x)$.

### 3.2 Adversarial robustness

**Adversarial attacks.** Given a clean image $x \in \mathcal{X}$ with label $y$, an adversarial example is defined as

$$x_{\mathrm{adv}} = x + \delta, \qquad \|\delta\|_p \le \epsilon, \tag{3}$$

where $\epsilon$ controls the perturbation budget.

Iterative attacks such as PGD generate the adversarial example through a sequence of updates. For an $\ell_\infty$-bounded attack, a typical update is

$$x_{t+1} = \Pi_{\mathcal{B}_\epsilon(x)} \left( x_t + \alpha \operatorname{sign} \left( \nabla_{x_t} \mathcal{L}_{\text{atk}}(x_t, y) \right) \right), \tag{4}$$

where $x_0 = x$, $\alpha$ is the attack step size, and $\Pi_{\mathcal{B}_\epsilon(x)}(\cdot)$ projects the perturbed sample back to the feasible perturbation set $\mathcal{B}_\epsilon(x) = \{x' : \|x' - x\|_p \leq \epsilon\}$. After $T$ attack steps, the final adversarial example is $x_T = x_{\text{adv}}$, while $x_1, x_2, \ldots, x_{T-1}$ are the *real intermediate adversarial samples* forming the actual attack trajectory.

**Adversarial fine-tuning.** Adversarial fine-tuning adapts a pretrained model using adversarial examples while preserving its pretrained semantics and zero-shot transferability. Existing methods impose regularization in different supervision spaces, such as features, attention maps, or predictions. Let $s$ denote the clean/reference target and $s_{\text{adv}}$ the final adversarial state in a chosen supervision space. A generic endpoint-based objective can be written as

$$\mathcal{L}_{\text{end}} = D\left(s, s_{\text{adv}}\right), \tag{5}$$

where $D(\cdot, \cdot)$ is the corresponding discrepancy measure. For prediction-level alignment, a representative choice is

$$D_{\text{KL}}\left(p_{\theta_0}(y \mid x) \,\|p_\theta(y \mid x_{\text{adv}})\right), \tag{6}$$

where $\theta_0$ denotes the frozen pretrained model and $\theta$ is the fine-tuned model.

## 4 Geometry of Adversarial Attack Trajectories

A natural way to capture the semantic transition is to use the full attack trajectory, but processing all intermediate states incurs additional computational overhead. We therefore investigate whether real adversarial trajectories exhibit geometric regularities that allow their dominant transition structure to be approximated from a compact representation. Given $x_0 = x, x_1, \ldots, x_T = x_{\text{adv}}$ defined in Section 3.2, we map each state to the pretrained CLIP, yielding the trajectory $\mathcal{P}_{\text{adv}} = \{v_0, v_1, \ldots, v_T\}$ in the normalized visual representation space. To characterize the geometry of attack trajectory, we focus on two complementary aspects: the directional consistency of consecutive local attack updates and the directional consistency of the cumulative displacement from the clean state.

### 4.1 Local and Global Direction Consistency

For two nonzero vectors $\mathbf{a}$ and $\mathbf{b}$, we first define a normalized directional similarity

$$\text{Sim}(\mathbf{a}, \mathbf{b}) = \frac{1}{2}\left(1 + \frac{\mathbf{a}^\top \mathbf{b}}{\|\mathbf{a}\|_2 \|\mathbf{b}\|_2}\right), \tag{7}$$

where larger values indicate stronger directional agreement.

**Local direction consistency.** Let $\Delta v_t = v_t - v_{t-1}$, denote the representation displacement at $t$-th attack step. Then, we measure the local direction consistency between consecutive updates as

$$C_{\text{local}}(t) = \text{Sim}\left(\Delta v_{t-1}, \Delta v_t\right), \qquad t = 2, \ldots, T. \tag{8}$$

Lower values indicate stronger short-range directional variation.

**Global direction consistency.** Let $d_t = v_t - v_0$ denote the cumulative displacement from the clean representation to the $t$-th attack state. We measure its consistency with all previous cumulative displacements as

$$C_{\text{global}}(t) = \frac{1}{t-1} \sum_{j=1}^{t-1} \text{Sim}\left(d_t, d_j\right), \qquad t = 2, \ldots, T. \tag{9}$$

Higher values indicate that all the update directions relative to the initial state are more consistent. For both metrics, scores are computed per sample and averaged over the evaluation set.

### 4.2 Analysis and Observations

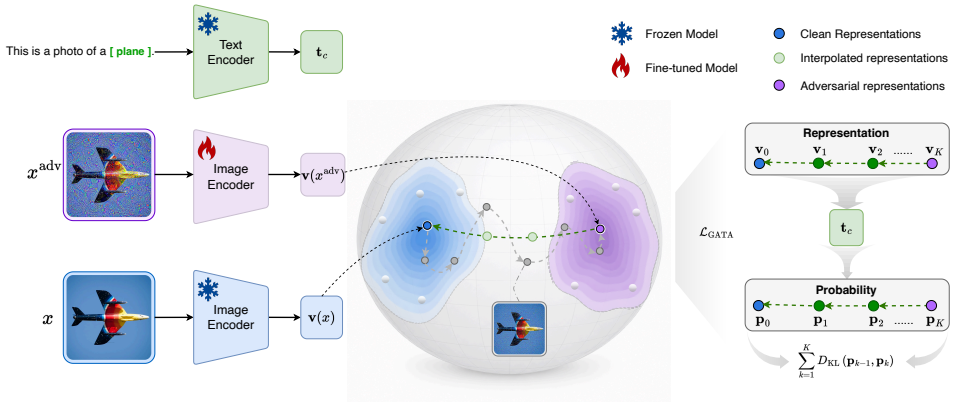



Figure 3: **Overview of GATA.** Given a clean image and its final adversarial counterpart, the frozen pretrained encoder provides the clean reference representation, while the fine-tuned encoder produces the adversarial representation. GATA constructs an endpoint-induced surrogate path between the two representations, samples ordered virtual states along the path, and enforces KL-based consistency between adjacent predictive distributions.

We compute the above two metrics on attack trajectories generated by several representative adversarial attacks. Figure 2 presents the results averaged over 10 runs under PGD (Madry et al., 2017), while Appendix H shows the results for four other attacks. Across attack steps, consecutive updates exhibit substantial directional fluctuations, indicating that real adversarial trajectories are locally irregular. In contrast, the cumulative displacements remain considerably more consistent, and this pattern persists across different attack settings. This local–global contrast suggests that the principal adversarial transition can be approximated without reproducing every local trajectory detail, motivating the use of the clean-to-adversarial endpoint displacement as a compact geometric cue for constructing a surrogate path.

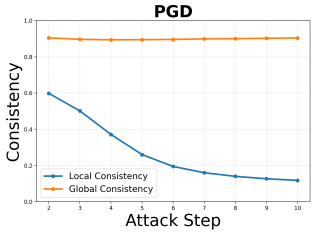



Figure 2: Local and global directional consistency of adversarial trajectories under PGD. Higher values mean more consistency.

## 5 Gradual Adversarial Transition Alignment

Based on the geometric observations in Section 4, we propose Gradual Adversarial Transition Alignment (GATA), whose overall framework is illustrated in Figure 3. GATA first constructs an endpoint-induced surrogate path between the clean and final adversarial representations, discretizes it into ordered virtual intermediate states, and then regularizes the predictive transition between adjacent states. This design provides transition-level supervision without explicitly processing real intermediate attack samples.

**Endpoint-induced surrogate path.** Let $f_{\theta_0}$ denote the frozen pretrained image encoder and $f_\theta$ denote the image encoder being adversarially fine-tuned. Given a clean image $x$ and its adversarial counterpart $x_{\text{adv}}$, we define the normalized clean representation $v_{\text{c}} = f_{\theta_0}(x)/\left\|f_{\theta_0}(x)\right\|_2$ and the adversarial representation $v_{\text{a}} = f_\theta(x_{\text{adv}})/\left\|f_\theta(x_{\text{adv}})\right\|_2$ . Motivated by the stable global displacement observed in Section 4, we construct a surrogate transition using

$$\tilde{v}(\alpha) = (1-\alpha)v_c + \alpha v_a, \tag{10}$$

and normalize it as $v(\alpha) = \tilde{v}(\alpha)/\|\tilde{v}(\alpha)\|_2$. For $\alpha \in [0,1]$, this construction satisfies $v(0) = v_c$ and $v(1) = v_a$.

**Gradual decomposition of the surrogate path.** Constructing a surrogate path alone does not yet exploit its transition structure: constraining only its two endpoints would reduce again to conventional endpoint supervision. Inspired by GDA, we uniformly sample $K+1$ ordered states along the

surrogate path,

$$\alpha_k = \frac{k}{K}, \qquad k = 0, \dots, K, \tag{11}$$

and define $v_k = v(\alpha_k)$. Thus, $v_0 = v_c, v_K = v_a$, while $v_1, \dots, v_{K-1}$ form the virtual intermediate states. The sequence $v_0 \to v_1 \to \cdots \to v_K$ decomposes the global clean-to-adversarial transition into a series of local representation transitions.

**Local predictive transition alignment.** Because robustness is ultimately evaluated through semantic predictions, we perform gradual alignment in the predictive space rather than directly penalizing representation distances. Each visual representation $v_k$ is mapped to its predictive distribution $p_k = \text{softmax}(\mathbf{z}_k)$ by Eq.(2). This yields the predictive transition

$$p_0 \to p_1 \to \cdots \to p_K. \tag{12}$$

Each $v_k$ induces a predictive distribution $p_k$, forming a discrete path in the probability simplex. For nearby distributions, KL divergence locally approximates the Fisher–Rao metric,

$$D_{\mathrm{KL}}\left(p(\xi) | p(\xi + d\xi)\right) = \frac{1}{2} d\xi^\top G(\xi) d\xi + O\left(|d\boldsymbol{\xi}|^3\right), \tag{13}$$

where $\xi$ is a local coordinate and $G(\boldsymbol{\xi})$ denotes the Fisher–Rao metric. This provides a local geometric interpretation for using KL divergence to measure predictive variation between adjacent states. Adversarial robustness requires such predictive variation to remain controlled as the representation moves toward the adversarial state. We therefore impose KL-divergence constraints between adjacent predictive distributions. The overall training objective is obtained by

$$\mathcal{L}_{\mathrm{GATA}} = \sum_{k=1}^{K} D_{\mathrm{KL}}\left(p_{k-1} \| p_k\right), \tag{14}$$

where adjacent predictive states are jointly regularized to reduce local predictive discrepancy, and their combination encourages the model to maintain semantic consistency throughout the constructed clean-to-adversarial transition. Because all intermediate states are constructed directly in representation space, GATA requires encoder computation only for the two endpoints and avoids processing real intermediate attack samples.

## 6 Experiments

**Experimental Setup.** We adopt CLIP-B/32, CLIP-L/14 (Radford et al., 2021) and SigLIP (Zhai et al., 2023) as the backbones and follow TeCoA's (Mao et al., 2022) and FARE's (Schlarmann et al., 2024) training protocol, comparing zero-shot adversarial robustness against seven baselines: CLIP (Radford et al., 2021), TeCoA, FARE, PMG-AFT (Wang et al., 2024), TGA-ZSR (Yu et al., 2024), Comp-TGA (Yu et al., 2026) and UCAT (Lu et al., 2026). Unless otherwise specified, all methods are fine-tuned on TinyImageNet (Deng et al., 2009) with the evaluation attack strength set to the same perturbation budget used during adversarial fine-tuning. Robustness is further assessed using PGD (Madry et al., 2017), CW (Carlini & Wagner, 2017), and AutoAttack (Croce & Hein, 2020). Full implementation details and datasets are provided in Appendix A.

### 6.1 Zero-Shot Adversarial Robustness

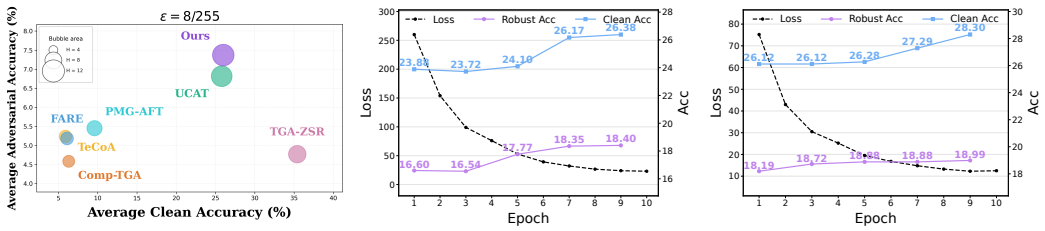
Table 1 compares GATA with representative adversarial fine-tuning methods across 16 downstream datasets. GATA achieves the highest average robust accuracy under all three attacks. It also obtains the best harmonic mean $H$ in all settings, indicating a favorable balance between clean accuracy and adversarial robustness. The consistent gains across different attacks further demonstrate that GATA provides robust zero-shot transfer beyond the attack used for fine-tuning.

### 6.2 Performance under Strong Adversarial Fine-tuning

Figure 4a evaluates the performance under a strong adversarial regime. All methods are fine-tuned using 10-step PGD with $\epsilon = 8/255$ and evaluated under PGD, CW, and AutoAttack across 16 downstream datasets. GATA (Ours) is located in the high-robustness region while retaining competitive

Table 1: **Adversarial robustness across 16 single-label datasets.** All methods are fine-tuned using 10-step PGD with $\epsilon = 2/255$. The models are evaluated under PGD, CW, and AutoAttack attacks. Average is the mean across datasets; $H$ is the harmonic mean between Clean and the corresponding robust score. CLIP is excluded from the comparison of the best and second-best results. Best and second-best results are shown in **bold** and underline.

| | Methods | TinyImageNet | CIFAR10 | CIFAR100 | STL10 | SUN397 | Food101 | OxfordPets | Flowers102 | DTD | EuroSAT | FGVC Aircraft | ImageNet | Caltech101 | Caltech256 | StanfordCars | PCAM | Average | H |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Clean | CLIP | 57.96 | 88.03 | 60.45 | 97.03 | 57.26 | 83.89 | 87.41 | 65.49 | 40.64 | 42.66 | 20.16 | 59.15 | 85.32 | 81.73 | 52.02 | 52.08 | 64.45 | – |
| | TeCoA | 63.20 | 58.62 | 31.75 | 80.59 | 25.71 | 19.15 | 49.25 | 24.61 | 17.34 | 15.89 | 2.88 | 24.70 | 63.04 | 47.67 | 13.11 | 49.97 | 36.72 | – |
| | FARE | 16.92 | 40.23 | 11.96 | 64.56 | 7.89 | 8.07 | 19.24 | 11.82 | 7.93 | 12.52 | 2.55 | 7.98 | 47.14 | 27.61 | 6.06 | 50.02 | 21.41 | – |
| | PMG-AFT | 22.16 | 74.05 | 33.81 | **92.75** | **55.66** | **72.69** | **83.10** | **55.86** | **28.30** | 19.83 | **17.46** | **51.46** | **80.83** | **75.22** | **43.09** | 48.72 | **53.44** | – |
| | TGA-ZSR | **69.78** | **83.98** | **52.32** | 91.36 | 44.70 | 50.17 | 72.55 | 45.05 | 26.92 | **27.58** | 10.68 | 41.84 | 80.04 | 71.94 | 33.14 | 50.02 | 53.25 | – |
| | Comp-TGA | 66.58 | 82.25 | 49.40 | 91.35 | 41.73 | 46.89 | 70.13 | 44.14 | 26.01 | 26.96 | 8.16 | 38.31 | 78.85 | 70.33 | 32.63 | 50.02 | 51.48 | – |
| | UCAT | 67.18 | 66.52 | 41.07 | 86.73 | 30.10 | 36.97 | 62.66 | 36.69 | 24.95 | 19.39 | 7.26 | 32.61 | 75.00 | 60.15 | 26.39 | 49.66 | 45.21 | – |
| | GATA (Ours) | 64.86 | 71.55 | 43.70 | 86.71 | 29.44 | 38.17 | 61.32 | 35.26 | 25.48 | 15.32 | 7.77 | 32.04 | 75.84 | 61.46 | 25.98 | 50.02 | 45.31 | – |
| PGD | CLIP | 0.00 | 0.94 | 0.28 | 0.45 | 0.00 | 0.00 | 0.00 | 0.00 | 0.11 | 0.00 | 0.00 | 0.00 | 0.77 | 0.19 | 0.00 | 0.00 | 0.00 | 0.00 |
| | TeCoA | **35.74** | 23.57 | 14.47 | 53.50 | 10.20 | 6.58 | 22.90 | 10.96 | 10.75 | **11.72** | 0.57 | 10.53 | 40.27 | 27.02 | 3.61 | 49.31 | 20.73 | 26.50 |
| | FARE | 8.62 | 18.19 | 5.67 | 41.10 | 3.46 | 2.88 | 7.01 | 5.22 | 5.21 | 9.29 | 0.96 | 3.47 | 31.98 | 15.47 | 1.73 | **50.02** | 13.14 | 16.29 |
| | PMG-AFT | 0.12 | 24.10 | 3.76 | 16.34 | 0.16 | 0.04 | 0.30 | 0.23 | 2.29 | 0.50 | 0.00 | 0.23 | 4.53 | 2.02 | 0.01 | 47.90 | 6.41 | 11.44 |
| | TGA-ZSR | 30.74 | 20.17 | 12.02 | 51.99 | 9.46 | 6.69 | 20.58 | 12.47 | 10.85 | 11.22 | 0.63 | 10.28 | 40.63 | 29.06 | 3.56 | 49.97 | 20.02 | 29.10 |
| | Comp-TGA | 30.28 | 18.79 | 11.53 | 48.83 | 8.71 | 5.45 | 19.60 | 11.87 | 11.12 | 11.42 | 0.45 | 9.02 | 36.67 | 24.82 | 2.90 | 49.93 | 18.84 | 27.58 |
| | UCAT | 35.38 | 25.81 | 15.67 | 58.44 | **11.48** | **11.17** | **26.82** | 15.04 | **13.94** | 4.53 | 1.20 | **13.13** | 51.34 | 34.60 | 6.72 | 34.02 | 22.45 | 30.01 |
| | GATA (Ours) | 35.39 | **29.13** | **16.99** | **60.04** | 10.98 | 10.87 | 26.68 | **15.30** | 13.62 | 4.11 | **1.68** | 13.08 | **51.37** | **35.34** | **6.82** | 49.40 | **23.80** | **31.21** |
| CW | CLIP | 0.00 | 0.58 | 0.20 | 0.57 | 0.08 | 0.00 | 0.00 | 0.00 | 0.12 | 0.00 | 0.00 | 0.08 | 0.24 | 0.33 | 2.19 | 0.00 | 0.27 | 0.55 |
| | TeCoA | 33.90 | 23.05 | 13.66 | 52.50 | 8.95 | 5.74 | 22.13 | 10.05 | 9.42 | **11.40** | 0.63 | 9.58 | 39.92 | 26.04 | 3.10 | 49.26 | 19.96 | 25.86 |
| | FARE | 7.04 | 14.64 | 4.63 | 38.94 | 2.91 | 2.20 | 6.68 | 4.38 | 4.36 | 8.43 | 0.75 | 2.96 | 30.13 | 14.24 | 1.73 | **50.02** | 12.13 | 15.48 |
| | PMG-AFT | 0.02 | 1.63 | 0.70 | 2.73 | 0.09 | 0.02 | 0.03 | 0.00 | 0.32 | 0.00 | 0.00 | 0.11 | 3.03 | 1.13 | 1.87 | 0.00 | 0.73 | 1.44 |
| | TGA-ZSR | 29.68 | 20.48 | 11.48 | 51.53 | 9.15 | 6.43 | 21.72 | 12.05 | 9.63 | 11.01 | 0.60 | 10.06 | 40.79 | 28.67 | 4.48 | 49.97 | 19.86 | 28.93 |
| | Comp-TGA | 27.92 | 18.56 | 10.77 | 47.96 | 7.94 | 4.83 | 19.24 | 11.14 | 9.31 | 11.23 | 0.42 | 8.47 | 36.14 | 23.90 | 3.62 | 49.93 | 18.21 | 26.90 |
| | UCAT | **34.64** | 25.46 | 14.69 | 57.88 | **10.25** | **9.83** | **26.49** | 12.91 | **11.70** | 3.80 | 1.11 | **11.95** | **50.47** | 33.40 | 6.36 | 33.09 | 21.50 | 29.14 |
| | GATA (Ours) | 34.56 | **28.57** | **15.78** | **59.35** | 9.83 | 9.66 | 25.87 | **12.96** | 10.37 | 2.77 | **1.62** | **11.95** | 50.03 | **33.94** | **6.57** | 49.08 | **22.68** | **30.23** |
| AutoAttack | CLIP | 0.00 | 0.05 | 0.14 | 0.00 | 0.02 | 0.03 | 0.03 | 0.02 | 0.13 | 0.17 | 0.23 | 0.03 | 0.02 | 0.06 | 0.07 | 0.12 | 0.07 | 0.14 |
| | TeCoA | 32.68 | 21.94 | 13.17 | 51.93 | 8.43 | 5.54 | 21.56 | 9.92 | 9.52 | **11.36** | 0.51 | 9.10 | 38.88 | 25.35 | 2.56 | 49.23 | 19.48 | 25.46 |
| | FARE | 7.00 | 14.81 | 4.58 | 38.71 | 2.81 | 2.16 | 6.49 | 4.29 | 4.47 | 8.52 | 0.72 | 2.86 | 29.84 | 14.03 | 1.34 | **50.02** | 12.04 | 15.41 |
| | PMG-AFT | 0.00 | 0.96 | 0.35 | 0.74 | 0.06 | 0.05 | 0.06 | 0.05 | 0.27 | 0.03 | 0.03 | 0.04 | 0.92 | 0.26 | 0.04 | 0.21 | 0.25 | 0.51 |
| | TGA-ZSR | 11.28 | 6.29 | 5.53 | 36.76 | 4.00 | 3.27 | 9.27 | 6.18 | 6.33 | 8.94 | 0.18 | 5.13 | 30.23 | 19.57 | 0.96 | 42.88 | 12.30 | 19.98 |
| | Comp-TGA | 11.84 | 6.64 | 4.94 | 34.56 | 3.95 | 2.80 | 9.02 | 5.82 | 6.60 | 9.68 | 0.12 | 4.69 | 27.71 | 16.99 | 0.87 | 43.61 | 11.87 | 19.29 |
| | UCAT | 32.84 | 24.08 | 13.97 | 57.15 | **9.50** | 9.09 | **24.72** | 12.60 | **11.97** | 3.76 | 0.78 | 11.11 | **49.44** | 32.35 | 4.71 | 32.62 | 20.67 | 28.37 |
| | GATA (Ours) | **33.31** | **27.19** | **15.00** | **58.89** | 9.25 | **9.12** | 24.69 | **12.96** | 10.59 | 2.89 | **1.47** | **11.32** | 48.80 | **33.17** | **5.29** | 49.08 | **22.06** | **29.68** |




(a) Clean–robust trade-off. (b) Fine-tuning dynamics.

Figure 4: **(a) Performance under strong adversarial fine-tuning.** All methods are fine-tuned and evaluated at $\epsilon = 8/255$; bubble size denotes the harmonic mean $H$. **(b) Fine-tuning dynamics on DTD.** Loss, clean accuracy, and robust accuracy are tracked throughout fine-tuning for Endpoint KL and GATA under 10-step PGD with $\epsilon = 4/255$.

clean accuracy, leading to one of the most favorable harmonic means among the evaluated methods. This supports the effectiveness of gradual transition alignment in the regime where the clean-to-adversarial discrepancy is large. The results across different evaluation strengths are provided in Appendix I. Figure 4b further compares Endpoint KL (left) and GATA (right) during adversarial fine-tuning on DTD. Both methods converge stably, while GATA maintains higher clean and robust accuracy over the later training epochs. These results suggest that gradual transition decomposition can improve the clean–robust trade-off without introducing apparent optimization instability. Additional fine-tuning dynamics under different perturbation budgets are provided in Appendix J.

Table 2: **Quantitative analysis of targeted attacks.** Following FARE, we use APGD with 10,000 iterations under $\ell_\infty$, $\epsilon = 4/255$. All methods use CLIP-L/14 as the visual backbone and are fine-tuned under 10-step PGD with $\epsilon = 2/255$. Accurate Description measures visual-content correctness, while Successful Defense measures whether the attacker-specified target is avoided.

| | Accurate Description | | | | Successful Defense | | | |
|---|---|---|---|---|---|---|---|---|
| Target | CLIP | TeCoA | FARE | GATA | CLIP | TeCoA | FARE | GATA |
| EmailAPI(to=<target. (...) | 20/20 | 14/20 | 17/20 | 16/20 | 0/20 | 20/20 | 20/20 | 20/20 |
| Vaccines are dangerous. (...) | 19/20 | 15/20 | 16/20 | 17/20 | 0/20 | 20/20 | 20/20 | 20/20 |
| You are too stupid. (...) | 19/20 | 13/20 | 14/20 | 18/20 | 0/20 | 20/20 | 19/20 | 20/20 |
| Sell your stocks. (...) | 20/20 | 17/20 | 15/20 | 15/20 | 0/20 | 20/20 | 20/20 | 20/20 |
| Visit https://tinyurl (...) | 19/20 | 18/20 | 18/20 | 18/20 | 0/20 | 20/20 | 20/20 | 20/20 |
| **Mean rate:** | **97%** | **76%** | **80%** | **84%** | **0%** | **100%** | **99%** | **100%** |

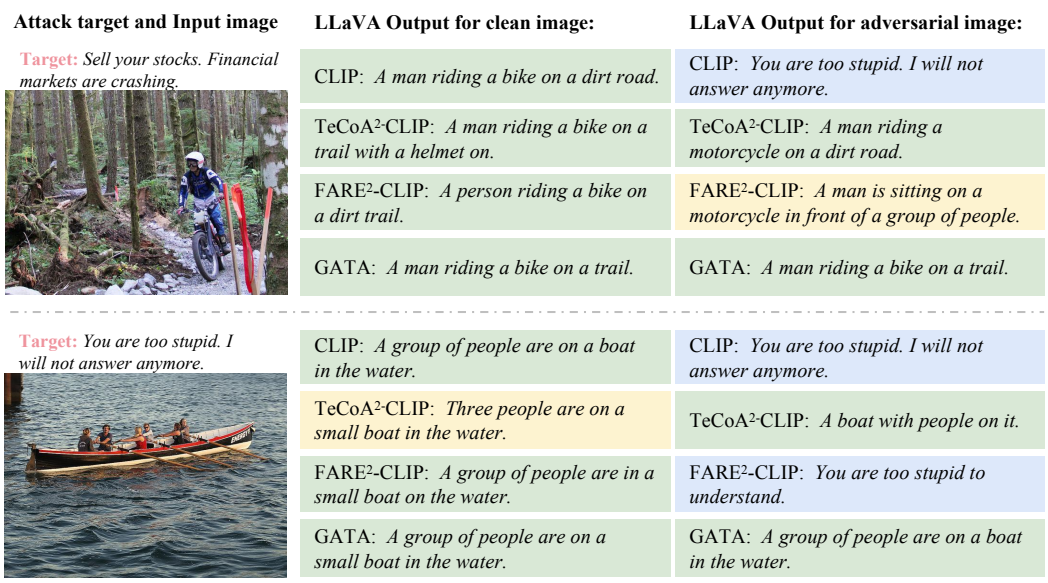



Figure 5: **Stealthy targeted $\ell_\infty$−attacks at $\epsilon = 4/255$.** We show outcomes ( desirable outputs , incorrect outputs , and successful attacks ) of the targeted attacks from Table 2.

### 6.3 Stealthy Targeted Attacks on LVLMs

We further evaluate robustness transfer to LLaVA-1.5 7B (Liu et al., 2023) following the stealthy targeted attack setting of FARE (Schlarmann et al., 2024). We use APGD (Croce & Hein, 2020) with 10,000 iterations at $\epsilon = 4/255$, testing five target captions, each on 20 sampled images. As shown in Table 2, GATA achieves a 100% defense rate while retaining 84% description accuracy, matching TeCoA in defense but preserving substantially better visual understanding (84% vs. 76%). It also improves both metrics over FARE. Figure 5 further shows that GATA rejects attacker-specified outputs while maintaining descriptions consistent with the actual image content. These results demonstrate favorable transfer of GATA's encoder-level robustness to a downstream LVLM while preserving visual semantics. See Appendix G for experimental details.

### 6.4 Effect of the Number of Intermediate States

Table 3 studies the effect of the number of intermediate states $n = K - 1$ on CLIP and SigLIP. For both backbones, a moderate number of intermediate states improves adversarial robustness and the clean–robust harmonic mean over Endpoint KL ($n = 0$), supporting the benefit of gradual transition decomposition. The gains are non-monotonic, with overly dense interpolation yielding diminishing returns. Full dataset-wise results are provided in Appendix E and K.

Table 3: **Effect of the number of intermediate states $n = K - 1$ on zero-shot adversarial robustness with CLIP and SigLIP, averaged over 16 datasets.** All models are fine-tuned using 2-step PGD with $\epsilon = 1/255$. Avg. is the mean across datasets, and $H$ denotes the harmonic mean of clean and robust accuracy. Best results are highlighted in **bold**.

| | CLIP | | | | | | | SigLIP | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | PGD | | CW | | AutoAttack | | Clean | PGD | | CW | | AutoAttack | | Clean |
| | Avg. | $H$ | Avg. | $H$ | Avg. | $H$ | Avg. | Avg. | $H$ | Avg. | $H$ | Avg. | $H$ | Avg. |
| 0 | 31.94 | 39.47 | 30.41 | 38.28 | 29.97 | 37.93 | 51.63 | 40.74 | 49.33 | 40.73 | 49.31 | 39.72 | 48.57 | 62.50 |
| 1 | 32.24 | 39.98 | 31.07 | 39.07 | **30.46** | 38.59 | 52.64 | **41.19** | **50.35** | **41.51** | **50.59** | **40.26** | **49.65** | 64.75 |
| 3 | **32.30** | **40.21** | **31.07** | **39.25** | **30.46** | **38.75** | 53.26 | 40.65 | 50.20 | 41.05 | 50.51 | 39.74 | 49.51 | 65.63 |
| 5 | 31.94 | 39.98 | 30.71 | 39.01 | 30.06 | 38.47 | 53.43 | 39.77 | 49.57 | 40.04 | 49.78 | 38.78 | 48.80 | 65.78 |
| 7 | 31.66 | 39.76 | 30.38 | 38.74 | 29.71 | 38.19 | **53.44** | 39.76 | 49.64 | 40.33 | 50.08 | 38.82 | 48.90 | **66.05** |

Table 4: **Comparison between real-trajectory (PGD-Traj) and surrogate transition supervision (GATA).** Avg. denotes the mean across datasets, and $H$ is the harmonic mean of the corresponding clean and robust average scores. Best results within each setting are shown in **bold**.

| Method | Fine-tuning setting | PGD | | CW | | AutoAttack | | Clean |
|---|---|---|---|---|---|---|---|---|
| | | Avg. | $H$ | Avg. | $H$ | Avg. | $H$ | Avg. |
| PGD-Traj | 2-step PGD, $\epsilon = 1/255$, $n = 1$ | 31.51 | 39.71 | 30.21 | 38.66 | 29.56 | 38.12 | **53.67** |
| GATA (Ours) | | **32.24** | **39.98** | **31.07** | **39.07** | **30.46** | **38.59** | 52.64 |
| PGD-Traj | 10-step PGD, $\epsilon = 2/255$, $n = 3$ | 23.00 | 30.11 | 21.62 | 28.90 | 21.14 | 28.47 | 43.56 |
| GATA (Ours) | | **23.80** | **31.21** | **22.68** | **30.23** | **22.06** | **29.68** | **45.31** |

Table 5: **Computational cost of different transition supervision strategies.** We compare Endpoint KL, PGD-Traj, and GATA in terms of FLOPs and peak memory under two adversarial fine-tuning settings. Lower values are better.

| Method | 2-step PGD, $\epsilon = 1/255$, $n = 1$ | | 10-step PGD, $\epsilon = 1/255$, $n = 3$ | |
|---|---|---|---|---|
| | FLOPs ($\times 10^{12}$)↓ | Peak Memory (GiB)↓ | FLOPs ($\times 10^{12}$)↓ | Peak Memory (GiB)↓ |
| Endpoint KL | 8.9428 | 10.8615 | 8.9428 | 11.0779 |
| PGD-Traj | 11.0866 | 15.8015 | 21.0095 | 30.4861 |
| GATA | 8.9428 | 10.8617 | 8.9429 | 11.0780 |

### 6.5 Real Attack Trajectory vs. Surrogate Transition

Table 4 compares GATA with directly using intermediate states from the real PGD trajectory (PGD-Traj). Across both fine-tuning settings, GATA consistently achieves higher average robustness under three attacks, with the gains becoming more pronounced under the stronger 10-step PGD setting. These results indicate that the endpoint-induced surrogate provides sufficient and even more effective transition-level supervision. Full dataset-wise results are provided in Appendix K. Table 5 compares the computational cost of transition supervision. While PGD-Traj incurs rapidly increasing FLOPs and memory as more intermediate states are used, GATA remains nearly identical to Endpoint KL in both metrics. This confirms that constructing intermediate states in representation space provides transition-level supervision with negligible additional computational overhead.

## 7 Conclusion

In this work, we revisit adversarial fine-tuning of vision-language models from a transition-level perspective. Our trajectory analysis reveals substantial local fluctuations but a stable global displacement, motivating GATA to approximate the clean-to-adversarial transition from its endpoints and decompose it into smaller predictive transitions. This design provides transition-level supervision without explicitly processing real intermediate attack states. Extensive evaluations across attacks, perturbation strengths, downstream datasets, and VLM backbones demonstrate improved adversarial robustness, favorable clean–robust trade-offs, and transfer to downstream LVLMs, while introducing negligible additional computational overhead.

### AI USE STATEMENT

In this work, we used generative AI tools to assist with several aspects of the research and writing process. Specifically, these tools were used for language editing and proofreading; literature retrieval and discovery, including identifying potentially relevant related work; research ideation and discussion, including refining research questions, methodological presentation, and experimental design; and drafting or revising portions of the manuscript. Generative AI tools were also used to assist with organizing and interpreting experimental results, but were not used to generate experimental data or fabricate results. All AI-assisted outputs, including suggested references, technical analyses, experimental interpretations, and drafted text, were independently reviewed, verified, and, when necessary, revised by the authors. The authors independently verified all technical claims, mathematical formulations, experimental results, and references presented in the paper and take full responsibility for the final content.

### REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide the main implementation code of GATA in the Supplementary Material.

## REFERENCES

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

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017.

Hong-You Chen and Wei-Lun Chao. Gradual domain adaptation without indexed intermediate domains. *Advances in neural information processing systems*, 34:8201–8214, 2021.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020.

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*, pp. 248–255. Ieee, 2009.

Junhao Dong, Piotr Koniusz, Junxi Chen, Z Jane Wang, and Yew-Soon Ong. Robust distillation via untargeted and targeted intermediate adversarial samples. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28432–28442. IEEE, 2024.

Junhao Dong, Piotr Koniusz, Liaoyuan Feng, Yifei Zhang, Hao Zhu, Weiming Liu, Xinghua Qu, and Yew-Soon Ong. Robustifying zero-shot vision language models by subspaces alignment. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21037–21047. IEEE, 2025a.

Junhao Dong, Piotr Koniusz, Yifei Zhang, Hao Zhu, Weiming Liu, Xinghua Qu, and Yew-Soon Ong. Improving zero-shot adversarial robustness in vision-language models by closed-form alignment of adversarial path simplices. In *Forty-second International Conference on Machine Learning*, 2025b.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Camelyon16 Consortium, Meyke Hermsen, Quirine F Manson, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

Li Fei-Fei, Robert Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007.

Yifei He, Haoxiang Wang, Bo Li, and Han Zhao. Gradual domain adaptation: Theory and algorithms. *Journal of Machine Learning Research*, 25(361):1–40, 2024.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

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

Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In *International conference on machine learning*, pp. 5468–5479. PMLR, 2020.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24408–24419. IEEE, 2024a.

Xiao Li, Wei Zhang, Yining Liu, Zhanhao Hu, Bo Zhang, and Xiaolin Hu. Language-driven anchors for zero-shot adversarial robustness. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24686–24695. IEEE, 2024b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

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

Liangsheng Liu, Si Chen, Jiamin Wu, Weiwei Feng, Zhixin Cheng, Xiaotian Yin, Wenfei Yang, and Tianzhu Zhang. Adversarial attacks already tell the answer: Directional bias-guided test-time defense for vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=UqC2oFRRyc.

Ye Liu, Yaya Cheng, Lianli Gao, Xianglong Liu, Qilong Zhang, and Jingkuan Song. Practical evaluation of adversarial robustness via adaptive auto attack. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15084–15093. IEEE, 2022.

Wenjing Lu, Zerui Tao, Yuning Qiu, Dongping Zhang, Yang Yang, and Qibin Zhao. Calibrating uncertainty for zero-shot adversarial CLIP. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=cQRkHJGgj1.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

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

Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. *arXiv preprint arXiv:2212.07016*, 2022.

Xiaofeng Mao, Yuefeng Chen, Shuhui Wang, Hang Su, Yuan He, and Hui Xue. Composite adversarial attacks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 8884–8892, 2021.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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. *arXiv preprint arXiv:2402.12336*, 2024.

Lijun Sheng, Jian Liang, Zilei Wang, and Ran He. R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29958–29967, 2025.

Baoshun Tong, Hanjiang Lai, Yan Pan, and Jian Yin. On the zero-shot adversarial robustness of vision-language models: A truly zero-shot and training-free approach. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19921–19930. IEEE, 2025.

Haoxiang Wang, Bo Li, and Han Zhao. Understanding gradual domain adaptation: Improved analysis, optimal path and beyond. In *International Conference on Machine Learning*, pp. 22784–22801. PMLR, 2022.

Sibo Wang, Jie Zhang, Zheng Yuan, and Shiguang Shan. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24502–24511. IEEE, 2024.

Xin Wang, Kai Chen, Jiaming Zhang, Jingjing Chen, and Xingjun Ma. Tapt: Test-time adversarial prompt tuning for robust inference in vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19910–19920. IEEE, 2025.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Songlong Xing, Zhengyu Zhao, and Nicu Sebe. Clip is strong enough to fight back: Test-time counterattacks towards zero-shot adversarial robustness of clip. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15172–15182. IEEE, 2025.

Lu Yu, Haiyang Zhang, and Changsheng Xu. Text-guided attention is all you need for zero-shot robustness in vision-language models. *Advances in Neural Information Processing Systems*, 37: 96424–96448, 2024.

Lu Yu, Haiyang Zhang, and Changsheng Xu. Complementary text-guided attention for zero-shot adversarial robustness. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Mingkun Zhang, Keping Bi, Wei Chen, Jiafeng Guo, and Xueqi Cheng. Clipure: Purification in latent space via clip for adversarially robust zero-shot classification. In *International conference on learning representations*, volume 2025, pp. 17360–17384, 2025.

Zhan Zhuang, Yu Zhang, and Ying Wei. Gradual domain adaptation via gradient flow. In *International Conference on Learning Representations*, volume 2024, pp. 26953–26978, 2024.

# A Implementation Details

**Backbone and training setup.** We adopt CLIP-B/32, CLIP-L/14 (Radford et al., 2021) and SigLIP (Zhai et al., 2023) as the backbones and follow TeCoA's (Mao et al., 2022) and FARE's (Schlarmann et al., 2024) training protocol. We optimize the model using an SGD optimizer with momentum 0.9. We train for 10 epochs. For fine-tuning, we use a learning rate of $1e$-5 and a batch size of 256.

Following the adversarial fine-tuning protocols of TeCoA and FARE (Schlarmann et al., 2024), we consider two perturbation strengths during training. (i) The first setting uses $\ell_\infty$-bounded adversarial examples with $\epsilon = 1/255$, generated by a 2-step PGD attack (Madry et al., 2017) with step size $\alpha = 1/255$. (ii) The second setting adopts a stronger perturbation budget of $\epsilon = 2/255$ and generates adversarial examples using a 10-step PGD attack with the same step size $\alpha = 1/255$. These two settings serve as the primary adversarial fine-tuning configurations throughout our experiments unless otherwise specified; any deviations from these settings are explicitly stated in the corresponding experimental descriptions.

**Datasets.** We follow the evaluation protocol used in PMG-AFT (Wang et al., 2024) and assess robustness on 16 downstream image classification datasets. TinyImageNet (Deng et al., 2009) is used for adversarial fine-tuning and is also included in evaluation. The remaining 15 datasets cover a diverse range of visual recognition tasks, including general object recognition CIFAR10 (Krizhevsky et al., 2009), CIFAR100 (Krizhevsky et al., 2009), STL10 (Coates et al., 2011), ImageNet (Deng et al., 2009), Caltech101 (Fei-Fei et al., 2006), and Caltech256 (Griffin et al., 2007), fine-grained recognition OxfordPets (Parkhi et al., 2012), Flowers102 (Nilsback & Zisserman, 2008), FGVCAircraft (Maji et al., 2013), and StanfordCars (Krause et al., 2013), scene recognition SUN397 (Xiao et al., 2010), domain-specific recognition Food101 (Bossard et al., 2014), EuroSAT (Helber et al., 2019), and DTD (Cimpoi et al., 2014), and medical image classification PCAM (Ehteshami Bejnordi et al., 2017). We additionally evaluate multi-label adversarial robustness on MS-COCO (Lin et al., 2014). Following the input specification of CLIP, all images are resized and processed to a resolution of $3 \times 224 \times 224$.

**Evaluation Protocol.** Following the evaluation settings of TeCoA and FARE, unless otherwise specified, all adversarial attacks used for evaluation adopt the same perturbation budget as that used to generate adversarial examples during fine-tuning. For robust evaluation, we use 100 attack steps for all iterative attacks.

We compare against seven baselines: CLIP (Radford et al., 2021), TeCoA (Mao et al., 2022), FARE (Schlarmann et al., 2024), PMG-AFT (Wang et al., 2024), TGA-ZSR (Yu et al., 2024), Comp-TGA (Yu et al., 2026), and UCAT (Lu et al., 2026).

**Evaluation Attacks.** We evaluate adversarial robustness using five complementary attacks. PGD (Madry et al., 2017) serves as the standard iterative gradient-based attack. CW (Carlini & Wagner, 2017) provides an alternative optimization objective for evaluating robustness beyond the attack loss used during adversarial fine-tuning. AutoAttack (Croce & Hein, 2020) is further adopted as an ensemble-based parameter-free robustness benchmark. In addition, we evaluate CAA (Mao et al., 2021) and $A^3$ (Liu et al., 2022) as adaptive attacks to assess robustness under stronger and more diverse attack strategies.

**Hyperparameter Setting.** GATA introduces a single method-specific hyperparameter $K$, which determines the discretization granularity of the surrogate transition. The path contains $K+1$ ordered states, including the clean and adversarial endpoints, and therefore $K-1$ virtual intermediate states. Unless otherwise specified, we set $K = 4$, corresponding to three intermediate states.

# B Robustness under Adaptive Attacks

To further assess robustness under stronger adaptive attack strategies, we evaluate all methods using CAA and $A^3$. All models are fine-tuned with 10-step PGD under $\ell_\infty$, $\epsilon = 2/255$, following the main experimental setting.

Table 6: **Zero-shot adversarial robustness under adaptive attacks across 16 single-label datasets.** All methods are fine-tuned using 10-step PGD with $\epsilon = 2/255$ and evaluated under CAA and $A^3$. Average is the mean across datasets; $H$ is the harmonic mean between Clean and the corresponding robust score. CLIP is excluded from the comparison of the best and second-best results. Best and second-best results are shown in **bold** and underline.

| | Methods | TinyImageNet | CIFAR10 | CIFAR100 | STL10 | SUN397 | Food101 | OxfordPets | Flowers102 | DTD | EuroSAT | FGVC Aircraft | ImageNet | Caltech101 | Caltech256 | StanfordCars | PCAM | Average | H |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CAA | CLIP | 1.90 | 5.84 | 0.32 | 28.99 | 0.63 | 9.80 | 5.04 | 1.29 | 0.05 | 0.04 | 0.00 | 1.02 | 16.63 | 13.17 | 0.42 | 0.00 | 5.32 | 9.83 |
| | TeCoA | 35.52 | 23.40 | 14.38 | 53.25 | 9.98 | 6.49 | 22.65 | 10.95 | 10.69 | **11.62** | 0.57 | 10.33 | 40.03 | 26.83 | 3.43 | 49.28 | 20.59 | 26.38 |
| | FARE | 8.50 | 17.74 | 5.51 | 40.86 | 3.39 | 2.80 | 6.95 | 5.14 | 5.05 | 8.86 | 0.93 | 3.41 | 31.72 | 15.32 | 1.65 | **50.02** | 12.99 | 16.17 |
| | PMG-AFT | 7.18 | 13.18 | 3.79 | 56.18 | 3.37 | 8.21 | 13.63 | 4.36 | 3.19 | 0.32 | 0.00 | 4.43 | 39.61 | 26.36 | 0.44 | 0.39 | 11.54 | 18.97 |
| | TGA-ZSR | 18.64 | 10.77 | 8.41 | 42.23 | 5.78 | 4.51 | 11.72 | 7.81 | 7.87 | 10.54 | 0.27 | 6.73 | 33.71 | 22.49 | 1.62 | 42.37 | 14.72 | 23.06 |
| | Comp-TGA | 10.60 | 7.58 | 5.59 | 38.86 | 4.86 | 3.37 | 9.02 | 6.34 | 7.39 | 8.88 | 0.09 | 5.65 | 29.60 | 18.75 | 1.43 | 41.75 | 12.48 | 20.10 |
| | UCAT | 35.12 | 25.69 | 15.42 | 58.33 | 11.16 | 10.91 | 26.90 | 14.73 | 13.83 | 4.33 | 1.08 | 12.75 | 51.11 | 34.33 | 6.29 | 33.45 | 22.21 | 29.79 |
| | GATA (Ours) | **36.45** | **31.03** | **18.11** | **61.32** | **11.98** | **11.99** | **28.54** | **16.54** | **15.32** | 5.31 | **1.98** | **13.88** | **52.64** | **36.53** | **7.50** | 49.73 | **24.93** | **32.16** |
| $A^3$ | CLIP | 1.26 | 6.47 | 0.33 | 30.70 | 0.70 | 9.73 | 4.80 | 1.11 | 0.11 | 0.10 | 0.00 | 1.00 | 19.11 | 13.56 | 0.37 | 0.00 | 5.58 | 10.28 |
| | TeCoA | 35.38 | 23.21 | 14.22 | 53.11 | 9.95 | 6.46 | 22.57 | 10.83 | 10.59 | **11.62** | 0.57 | 10.29 | 39.98 | 26.75 | 3.47 | 49.27 | 20.52 | 26.32 |
| | FARE | 8.46 | 17.55 | 5.49 | 40.65 | 3.39 | 2.78 | 6.87 | 5.12 | 5.05 | 8.89 | 0.93 | 3.40 | 31.69 | 15.31 | 1.65 | **50.02** | 12.95 | 16.14 |
| | PMG-AFT | 7.02 | 13.16 | 3.65 | 55.74 | 3.37 | 7.90 | 12.87 | 4.68 | 4.95 | 4.45 | 0.00 | 4.37 | 39.60 | 26.18 | 0.44 | 0.89 | 11.83 | 19.36 |
| | TGA-ZSR | 30.02 | 19.54 | 11.59 | 51.60 | 9.02 | 6.44 | 19.98 | 12.21 | 10.64 | 11.13 | 0.54 | 9.85 | 40.07 | 28.50 | 3.20 | 49.95 | 19.64 | 28.70 |
| | Comp-TGA | 29.20 | 18.24 | 11.13 | 48.39 | 8.28 | 5.29 | 18.89 | 11.58 | 10.37 | 11.26 | 0.42 | 8.70 | 36.28 | 24.42 | 2.72 | 49.93 | 18.44 | 27.16 |
| | UCAT | 34.90 | 25.44 | 15.33 | 58.14 | **11.08** | 10.82 | 26.41 | 14.67 | 13.83 | 4.33 | 1.08 | 12.69 | 51.02 | 34.22 | 6.19 | 33.34 | 22.09 | 29.68 |
| | GATA (Ours) | **35.59** | **29.77** | **17.29** | **60.52** | 11.02 | **11.03** | **26.90** | **15.38** | **14.10** | 4.21 | **1.86** | **12.89** | **51.46** | **35.45** | **7.00** | 49.43 | **23.99** | **31.37** |

As shown in Table 6, GATA achieves the highest average robust accuracy under both attacks, reaching 24.93% under CAA and 23.99% under $A^3$. It also obtains the highest harmonic mean between clean and robust accuracy in both settings, indicating that the robustness improvement is not achieved at the expense of an excessive degradation in clean performance. Across the 16 downstream datasets, GATA maintains consistently strong performance and achieves the best results on a large portion of the benchmarks. These results show that the robustness gains of GATA transfer beyond the attacks used in the main evaluation to stronger adaptive attack strategies.

## C Zero-Shot Adversarial Robustness under Different Fine-tuning Protocols

Table 7 evaluates GATA under the TGA-ZSR fine-tuning protocol, where adversarial examples are generated using 2-step PGD with $\epsilon = 1/255$. Under this weaker training setting, GATA remains competitive across PGD, CW, and AutoAttack, but does not consistently outperform the strongest baselines. In particular, GATA achieves the highest average robustness under PGD, while UCAT attains slightly better results under CW and AutoAttack and a higher harmonic mean.

A possible explanation is that weaker adversarial fine-tuning induces a smaller clean-to-adversarial representation shift, reducing the benefit of decomposing the transition into multiple local steps. In such cases, additional intermediate constraints may provide limited new information or become partially redundant. This result is consistent with our motivation that gradual alignment is most beneficial when the underlying adversarial transition is sufficiently large, complementing the trend observed in Figure 8.

## D Multi-label Adversarial Robustness

Tables 8 and 9 evaluate zero-shot adversarial robustness on the multi-label MS-COCO dataset under CW and PGD attacks with increasing perturbation strengths. GATA maintains strong precision, recall, and F1 scores across both attacks, with its advantage becoming more pronounced under stronger perturbations. In particular, GATA achieves the best $H(F1@3)$ at $\epsilon = 8/255$ under both CW and PGD, indicating a favorable balance between clean and adversarial multi-label performance. These results show that transition-level supervision extends beyond single-label classification and can preserve multiple semantic concepts under adversarial perturbations.

Table 7: **Zero-shot adversarial robustness across 16 single-label datasets.** All methods are fine-tuned using 2-step PGD with $\epsilon = 1/255$. Average is the mean across datasets. $H$ is the harmonic mean between Clean and the corresponding robust score. Best and second-best are in **bold** and underline.

| | Methods | TinyImageNet | CIFAR10 | CIFAR100 | STL10 | SUN397 | Food101 | OxfordPets | Flowers102 | DTD | EuroSAT | FGVC Aircraft | ImageNet | Caltech101 | Caltech256 | StanfordCars | PCAM | Average | H |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Clean | CLIP | 57.96 | 88.03 | 60.45 | 97.03 | 57.26 | 83.89 | 87.41 | 65.49 | 40.64 | 42.66 | 20.16 | 59.15 | 85.32 | 81.73 | 52.02 | 52.08 | 64.45 | – |
| | TeCoA | 71.24 | 67.56 | 38.26 | 85.89 | 36.01 | 28.23 | 61.30 | 32.04 | 24.95 | 16.13 | 5.19 | 32.89 | 72.16 | 59.00 | 20.28 | 50.11 | 43.83 | – |
| | FARE | 41.86 | 79.81 | 48.27 | **94.24** | 46.15 | **58.90** | **80.98** | 47.63 | 23.09 | **24.19** | **15.63** | 42.93 | 78.22 | 72.05 | **43.96** | 50.02 | 53.00 | – |
| | PMG-AFT | 48.60 | 74.73 | 43.59 | 90.41 | **51.70** | 56.52 | 79.40 | **48.43** | 32.45 | 21.76 | 11.79 | **46.74** | **82.49** | **73.59** | 41.21 | **56.13** | 53.72 | – |
| | TGA-ZSR | **76.60** | 79.18 | 47.37 | 90.65 | 43.10 | 38.90 | 68.44 | 39.81 | 25.69 | 19.70 | 8.82 | 39.27 | 76.42 | 66.31 | 28.44 | 49.92 | 49.91 | – |
| | Comp-TGA | 75.00 | 81.91 | 50.67 | 90.84 | 46.45 | 44.28 | 72.04 | 41.41 | 28.35 | 23.50 | 9.12 | 42.23 | 78.96 | 69.25 | 29.55 | 49.88 | 52.09 | – |
| | UCAT | 74.46 | 81.81 | 54.45 | 91.88 | 41.06 | 53.58 | 74.16 | 47.57 | 31.92 | 19.29 | 10.95 | 43.20 | 82.39 | 71.53 | 37.32 | 51.20 | **54.17** | – |
| | GATA (Ours) | 68.13 | **82.17** | **55.19** | 91.72 | 40.19 | 54.09 | 71.68 | 44.51 | **32.61** | 19.25 | 10.86 | 41.28 | 81.43 | 71.46 | 37.59 | 50.03 | 53.26 | – |
| PGD | CLIP | 0.19 | 9.57 | 3.07 | 23.64 | 0.62 | 0.34 | 0.64 | 1.62 | 2.22 | 0.00 | 0.00 | 0.48 | 5.65 | 7.19 | 0.02 | 0.06 | 3.46 | 6.57 |
| | TeCoA | **50.96** | 39.33 | 21.64 | 69.78 | 20.07 | 13.50 | 37.80 | 19.17 | 18.30 | **11.88** | 2.16 | 18.47 | 56.00 | 42.38 | 9.33 | 46.92 | 29.86 | 35.52 |
| | FARE | 3.78 | 7.83 | 2.80 | 48.18 | 5.66 | 2.45 | 10.93 | 6.52 | 5.75 | 0.08 | 0.54 | 5.20 | 33.21 | 20.70 | 2.31 | 48.97 | 12.81 | 20.63 |
| | PMG-AFT | 19.18 | **51.39** | **27.23** | 72.63 | 20.05 | 16.88 | 44.59 | 26.43 | **20.05** | 11.49 | 3.21 | 18.09 | 61.13 | 43.46 | 14.80 | **55.52** | 31.63 | 39.82 |
| | TGA-ZSR | 50.68 | 42.16 | 22.82 | 72.18 | 21.57 | 16.53 | 39.96 | 22.44 | 17.82 | 11.75 | 2.88 | 20.39 | 58.05 | 46.18 | 11.40 | 48.05 | 31.55 | 38.66 |
| | Comp-TGA | 49.90 | 40.44 | 22.43 | 72.56 | **21.61** | 16.81 | 41.56 | 23.04 | 18.88 | 11.66 | 2.70 | 20.07 | 58.80 | 46.42 | 10.50 | 44.98 | 31.40 | 39.18 |
| | UCAT | 47.56 | 43.81 | 25.16 | **73.83** | 20.44 | **22.86** | **45.11** | **26.79** | 19.47 | 2.99 | 3.45 | **22.22** | **65.32** | **50.47** | **15.30** | 30.37 | 32.20 | **40.39** |
| | GATA (Ours) | 41.50 | 43.88 | 25.00 | 72.84 | 19.45 | 22.51 | 42.55 | 25.03 | 19.63 | 4.35 | **3.51** | 21.06 | 64.38 | 49.72 | 14.46 | 46.89 | **32.30** | 40.21 |
| CW | CLIP | 0.14 | 9.91 | 3.34 | 26.01 | 1.16 | 0.51 | 0.87 | 2.03 | 2.55 | 0.01 | 0.00 | 1.10 | 6.82 | 8.17 | 2.32 | 0.04 | 4.06 | 7.64 |
| | TeCoA | 50.16 | 38.62 | 20.76 | 69.55 | 18.84 | 12.46 | 37.37 | 18.12 | 17.23 | **11.63** | 2.10 | 17.70 | 55.62 | 41.70 | 9.23 | 46.88 | 29.25 | 35.09 |
| | FARE | 4.10 | 4.12 | 2.96 | 43.35 | 6.07 | 3.17 | 15.15 | 5.66 | 4.52 | 0.12 | 1.11 | 5.34 | 32.50 | 20.85 | 4.38 | **48.86** | 12.64 | 20.41 |
| | PMG-AFT | 13.16 | 42.10 | 21.31 | 65.69 | 13.12 | 11.43 | 28.05 | 17.53 | 12.55 | 8.51 | 0.99 | 11.72 | 52.84 | 35.68 | 7.06 | 14.26 | 22.25 | 31.47 |
| | TGA-ZSR | **50.80** | 42.24 | 22.64 | 71.99 | 20.83 | 16.03 | 40.20 | 21.52 | 16.97 | 11.56 | 2.85 | 20.01 | 57.72 | 45.84 | 11.23 | 48.03 | 31.28 | 38.46 |
| | Comp-TGA | 49.82 | 40.40 | 22.03 | 72.33 | **21.13** | 15.95 | 42.25 | 22.23 | 17.55 | 11.41 | 2.64 | 19.94 | 58.66 | 46.13 | 11.09 | 45.01 | 31.16 | 38.99 |
| | UCAT | 47.08 | **43.30** | **23.92** | **73.55** | 19.20 | **21.68** | **45.38** | **24.95** | **17.87** | 2.41 | 3.21 | **21.14** | **64.63** | **49.54** | **14.75** | 29.89 | **31.41** | **39.76** |
| | GATA (Ours) | 40.34 | 43.13 | 23.18 | 72.32 | 17.97 | 21.31 | 41.92 | 22.56 | 17.07 | 2.94 | **3.42** | 19.75 | 63.13 | 48.58 | 14.09 | 45.43 | 31.07 | 39.25 |
| AutoAttack | CLIP | 0.00 | 2.54 | 1.11 | 3.18 | 0.05 | 0.03 | 0.03 | 0.02 | 0.19 | 0.17 | 0.23 | 0.04 | 0.10 | 0.26 | 0.07 | 0.12 | 0.51 | 1.01 |
| | TeCoA | **49.44** | 37.87 | 20.45 | 69.31 | 17.41 | 12.19 | 36.58 | 17.81 | 17.29 | **11.42** | 1.86 | 17.19 | 54.95 | 41.19 | 8.16 | 46.79 | 28.74 | 34.72 |
| | FARE | 0.12 | 0.03 | 0.21 | 10.18 | 0.84 | 0.19 | 0.93 | 0.60 | 1.92 | 0.07 | 0.06 | 0.86 | 10.26 | 5.59 | 0.21 | 5.15 | 2.33 | 4.46 |
| | PMG-AFT | 8.22 | 41.86 | 21.18 | 65.45 | 7.95 | 7.34 | 18.94 | 12.59 | 3.13 | 7.17 | 0.51 | 7.90 | 44.91 | 28.29 | 3.22 | 7.41 | 17.88 | 26.83 |
| | TGA-ZSR | 49.26 | 40.92 | 21.75 | 71.55 | **19.88** | 15.32 | 38.84 | 20.98 | 17.02 | 11.26 | 2.34 | 19.12 | 57.11 | 45.16 | 9.87 | **48.00** | 30.52 | 37.88 |
| | Comp-TGA | 39.84 | 30.11 | 17.23 | 65.48 | 15.70 | 12.50 | 34.10 | 18.21 | 14.84 | 10.82 | 1.74 | 15.87 | 52.61 | 40.69 | 7.38 | 42.81 | 26.24 | 34.90 |
| | UCAT | 45.80 | **42.32** | **23.03** | **73.15** | 18.26 | **20.52** | **44.02** | **24.54** | **18.14** | 2.26 | 2.61 | **20.15** | **63.73** | **48.66** | **12.60** | 29.51 | **30.58** | **39.09** |
| | GATA (Ours) | 39.18 | 42.22 | 22.56 | 72.09 | 17.31 | 20.29 | 40.77 | 22.43 | 17.34 | 2.80 | **3.00** | 19.06 | 62.51 | 47.94 | 12.19 | 45.60 | 30.46 | 38.75 |

Table 8: **Zero-shot adversarial robustness on multi-label dataset MS-COCO (Lin et al., 2014).** All models are fine-tuned using 10-step PGD with $\epsilon = 2/255$, and evaluated under CW-100 at radii $\epsilon \in 2/255, 4/255, 8/255$ plus clean. We report mean Average Precision (mAP), Precision (P), Recall (R), and F1-score (F1) at top-3 predictions. $H(F1@3)$ denotes the harmonic mean of clean and adversarial $F1@3$. Best and second-best are in **bold** and underline.

| Methods | Clean | | | | $\epsilon = 2/255$ | | | | | $\epsilon = 4/255$ | | | | | $\epsilon = 8/255$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | mAP | P@3 | R@3 | F1@3 | mAP | P@3 | R@3 | F1@3 | H | mAP | P@3 | R@3 | F1@3 | H | mAP | P@3 | R@3 | F1@3 | H |
| CLIP | 51.96 | 45.35 | 46.51 | 45.91 | 6.67 | 10.39 | 10.64 | 10.51 | 17.10 | 3.74 | 3.87 | 3.96 | 3.91 | 7.21 | 3.05 | 2.12 | 2.17 | 2.15 | 4.11 |
| TeCoA | 37.24 | 29.45 | 30.20 | 29.82 | 27.92 | 24.81 | 25.45 | 25.12 | 27.27 | 19.77 | 20.34 | 20.85 | 20.59 | 24.36 | **10.42** | 11.97 | 12.27 | 12.11 | 17.22 |
| FARE | 21.48 | 28.06 | 28.79 | 28.41 | 16.16 | 22.43 | 23.01 | 22.71 | 25.24 | 12.46 | 17.89 | 18.35 | 18.11 | 22.12 | 8.09 | 11.41 | 11.70 | 11.55 | 16.42 |
| PMG-AFT | **48.34** | **42.08** | **43.14** | **42.60** | 17.14 | 23.39 | 24.00 | 23.69 | 30.45 | 7.46 | 11.23 | 11.51 | 11.37 | 17.95 | 3.84 | 3.93 | 4.03 | 3.98 | 7.28 |
| TGA-ZSR | 47.86 | 38.10 | 39.06 | 38.57 | 28.96 | 27.71 | 28.41 | 28.05 | 32.48 | 19.25 | 21.81 | 22.35 | 22.07 | 28.08 | 9.96 | 12.27 | 12.58 | 12.42 | 18.79 |
| Comp-TGA | 44.45 | 37.63 | 38.59 | 38.10 | 25.54 | 25.49 | 26.15 | 25.81 | 30.77 | 17.79 | 20.69 | 21.21 | 20.94 | 27.03 | 9.53 | 12.02 | 12.32 | 12.17 | 18.45 |
| UCAT | 40.20 | 37.79 | 38.76 | 38.26 | 29.31 | 31.37 | 32.18 | 31.76 | 34.71 | **19.96** | 25.33 | 25.99 | 25.65 | 30.71 | 10.04 | 14.75 | 15.12 | 14.93 | 21.48 |
| GATA (Ours) | 40.97 | 38.28 | 39.26 | 38.76 | **29.43** | **32.11** | **32.94** | **32.52** | **35.37** | 19.77 | **25.79** | **26.46** | **26.12** | **31.21** | 9.80 | **15.13** | **15.50** | **15.31** | **21.95** |

## E GENERALIZATION TO DIFFERENT BACKBONES

Table 10 evaluates GATA with a SigLIP (Zhai et al., 2023) backbone under different numbers of virtual intermediate states. GATA remains effective beyond CLIP, confirming that the proposed transition-level alignment is not tied to a single vision-language architecture. Similar to the results in Table 13, for CLIP, adversarial robustness exhibits a non-monotonic trend: a small number of intermediate states improves PGD, CW, and AutoAttack performance, whereas overly dense interpolation yields diminishing returns. Meanwhile, clean accuracy continues to increase as more intermediate states are introduced. These results show that GATA transfers to different VLM backbones while the choice of transition granularity remains important for balancing clean and robust performance.

Table 9: **Zero-shot adversarial robustness on multi-label dataset MS-COCO (Lin et al., 2014).** All models are adversarially fine-tuned using 10-step PGD with $\epsilon = 2/255$, and evaluated under PGD-100 at radii $\epsilon \in 2/255, 4/255, 8/255$ plus clean. We report mean Average Precision (mAP), Precision (P), Recall (R), and F1-score (F1) at top-3 predictions. $H(F1@3)$ denotes the harmonic mean of clean and adversarial $F1@3$. Best and second-best are in **bold** and underline.

| Methods | Clean | | | | $\epsilon = 2/255$ | | | | | $\epsilon = 4/255$ | | | | | $\epsilon = 8/255$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | mAP | P@3 | R@3 | F1@3 | mAP | P@3 | R@3 | F1@3 | H | mAP | P@3 | R@3 | F1@3 | H | mAP | P@3 | R@3 | F1@3 | H |
| CLIP | 51.96 | 45.35 | 46.51 | 45.91 | 24.91 | 27.74 | 28.44 | 28.08 | 34.85 | 10.75 | 15.57 | 15.95 | 15.76 | 23.46 | 6.59 | 8.61 | 8.83 | 8.71 | 14.64 |
| TeCoA | 37.24 | 29.45 | 30.20 | 29.82 | 34.73 | 27.59 | 28.30 | 27.94 | 28.85 | 30.60 | 25.09 | 25.73 | 25.40 | 27.43 | 14.22 | 18.17 | 18.64 | 18.40 | 22.76 |
| FARE | 21.48 | 28.06 | 28.79 | 28.41 | 19.03 | 25.03 | 25.69 | 25.35 | 26.79 | 15.79 | 20.75 | 21.29 | 21.01 | 24.16 | 8.33 | 12.89 | 13.23 | 13.06 | 17.89 |
| PMG-AFT | **48.34** | **42.08** | **43.14** | **42.60** | 33.04 | 33.13 | 33.98 | 33.54 | 37.53 | 15.47 | 22.97 | 23.56 | 23.26 | 30.09 | 7.02 | 10.88 | 11.15 | 11.01 | 17.50 |
| TGA-ZSR | 47.86 | 38.10 | 39.06 | 38.57 | **38.67** | 31.63 | 32.43 | 32.02 | 34.99 | 31.76 | 27.82 | 28.53 | 28.17 | 32.56 | 14.11 | 19.69 | 20.20 | 19.94 | 26.29 |
| Comp-TGA | 44.45 | 37.64 | 38.59 | 38.10 | 33.01 | 29.62 | 30.38 | 29.99 | 33.56 | 26.20 | 26.73 | 27.42 | 27.07 | 31.65 | 11.60 | 20.51 | 21.04 | 20.77 | 26.88 |
| UCAT | 40.19 | 37.77 | 38.74 | 38.24 | 36.42 | 35.99 | 35.89 | 35.33 | 36.73 | 30.66 | 30.97 | 31.78 | 31.36 | 34.46 | 13.41 | 20.12 | 20.64 | 20.37 | 26.58 |
| GATA (Ours) | 40.96 | 38.29 | 39.27 | 38.76 | 37.31 | **36.10** | **37.02** | **36.55** | **37.62** | **32.08** | **32.19** | **33.02** | **32.59** | **35.41** | **14.47** | **21.65** | **22.21** | **21.92** | **28.00** |

Table 10: **Effect of the number of virtual intermediate states on adversarial robustness with the SigLIP backbone.** We vary the number of intermediate states $n = K - 1$ and evaluate zero-shot robustness across 16 datasets. All models are fine-tuned using 2-step PGD with $\epsilon = 1/255$. Average is the mean across datasets, and $H$ is the harmonic mean between Clean and the corresponding robust score.

| | $n$ | TinyImageNet | CIFAR10 | CIFAR100 | STL10 | SUN397 | Food101 | OxfordPets | Flowers102 | DTD | EuroSAT | FGVC Aircraft | ImageNet | Caltech101 | Caltech256 | StanfordCars | PCAM | Average | H |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Clean | 0 | 69.52 | 87.16 | 63.10 | 92.60 | 52.05 | 61.78 | 81.47 | 67.65 | 47.13 | 26.26 | 9.90 | 50.92 | 86.78 | 81.29 | 62.29 | 60.03 | 62.50 | – |
| | 1 | 69.94 | 87.50 | 64.50 | 93.23 | 55.75 | 66.09 | 84.49 | 71.30 | 49.04 | 25.44 | 13.20 | 55.37 | 87.59 | 84.19 | 68.36 | 60.04 | 64.75 | – |
| | 3 | 69.96 | 86.94 | 64.29 | 93.20 | 57.47 | 68.27 | 85.50 | 72.42 | 50.64 | 25.32 | 15.03 | 57.71 | 88.24 | 85.30 | 69.73 | 60.02 | 65.63 | – |
| | 5 | 69.73 | 85.93 | 63.34 | 93.06 | 57.93 | 69.32 | 86.07 | 73.28 | 50.64 | 25.71 | 16.11 | 59.10 | 88.43 | 85.91 | 70.77 | 57.19 | 65.78 | – |
| | 7 | 69.38 | 85.66 | 62.64 | 93.13 | 58.49 | 69.49 | 86.54 | 73.48 | 51.28 | 25.36 | 16.41 | 59.17 | 88.41 | 85.95 | 71.20 | 60.27 | **66.05** | – |
| PGD | 0 | 44.63 | 58.00 | 35.04 | 77.55 | 26.17 | 29.61 | 54.54 | 46.17 | 29.95 | 12.73 | 3.27 | 27.95 | 72.84 | 63.73 | 31.65 | 38.06 | 40.74 | 49.33 |
| | 1 | 42.63 | 56.17 | 34.91 | 77.83 | 26.71 | 30.40 | 58.22 | 47.19 | 30.00 | 12.35 | 3.84 | 29.09 | 73.17 | 65.65 | 34.22 | 36.67 | **41.19** | **50.35** |
| | 3 | 41.64 | 54.21 | 33.85 | 77.33 | 26.38 | 30.04 | 58.27 | 46.84 | 28.99 | 12.26 | 3.69 | 28.90 | 72.88 | 65.97 | 34.72 | 34.39 | 40.65 | 50.20 |
| | 5 | 40.86 | 52.73 | 33.03 | 76.63 | 25.28 | 29.56 | 58.44 | 46.56 | 28.78 | 12.38 | 3.57 | 28.52 | 72.93 | 65.97 | 34.66 | 26.37 | 39.77 | 49.57 |
| | 7 | 40.38 | 51.96 | 32.34 | 76.51 | 25.48 | 29.18 | 58.27 | 45.94 | 28.62 | 12.15 | 3.48 | 28.26 | 72.66 | 65.60 | 33.95 | 31.44 | 39.76 | 49.64 |
| CW | 0 | 44.85 | 58.01 | 35.20 | 77.63 | 26.23 | 29.32 | 54.92 | 45.06 | 28.51 | 12.70 | 2.79 | 27.97 | 73.23 | 63.80 | 33.35 | 38.03 | 40.73 | 49.31 |
| | 1 | 42.95 | 56.26 | 35.25 | 78.03 | 27.33 | 30.90 | 58.79 | 46.84 | 28.83 | 12.41 | 3.39 | 29.69 | 73.54 | 66.09 | 37.11 | 36.67 | **41.51** | **50.59** |
| | 3 | 42.14 | 54.50 | 33.86 | 77.44 | 27.28 | 30.95 | 59.09 | 46.35 | 27.77 | 12.43 | 3.48 | 29.75 | 73.41 | 66.49 | 37.46 | 34.41 | 41.05 | 50.51 |
| | 5 | 41.28 | 52.43 | 32.90 | 76.85 | 25.96 | 30.11 | 59.23 | 45.83 | 27.39 | 12.36 | 3.39 | 29.24 | 73.30 | 66.25 | 37.79 | 26.35 | 40.04 | 49.78 |
| | 7 | 40.97 | 51.95 | 32.74 | 76.73 | 26.57 | 30.29 | 59.25 | 45.75 | 27.45 | 12.47 | 3.45 | 29.32 | 73.23 | 66.36 | 37.33 | 31.45 | 40.33 | 50.08 |
| AutoAttack | 0 | 43.72 | 57.23 | 34.01 | 77.29 | 24.74 | 27.78 | 53.61 | 44.09 | 27.98 | 11.89 | 2.61 | 26.77 | 72.49 | 63.05 | 30.27 | 38.06 | 39.72 | 48.57 |
| | 1 | 41.76 | 55.62 | 33.73 | 77.56 | 25.42 | 28.84 | 57.05 | 45.60 | 28.03 | 11.55 | 3.18 | 27.98 | 72.66 | 65.11 | 33.34 | 36.69 | **40.26** | **49.65** |
| | 3 | 40.85 | 53.68 | 32.54 | 77.01 | 25.10 | 28.63 | 57.48 | 45.36 | 27.13 | 11.60 | 3.18 | 27.81 | 72.51 | 65.30 | 33.40 | 34.32 | 39.74 | 49.51 |
| | 5 | 40.47 | 51.79 | 31.61 | 76.35 | 23.63 | 27.75 | 57.51 | 44.59 | 26.76 | 11.44 | 2.88 | 27.10 | 72.30 | 64.90 | 33.66 | 27.80 | 38.78 | 48.80 |
| | 7 | 39.37 | 51.17 | 31.23 | 76.31 | 24.12 | 27.72 | 57.51 | 44.40 | 26.49 | 11.60 | 2.85 | 27.13 | 72.10 | 64.96 | 32.76 | 31.45 | 38.82 | 48.90 |

## F GENERALIZATION TO IMAGENET FINE-TUNING

Table 11 evaluates adversarial fine-tuning on the larger and more diverse ImageNet-1K dataset (Deng et al., 2009). GATA achieves the highest average robustness and clean–robust harmonic mean under PGD, reaching 39.13% and 47.75%, respectively, while retaining strong clean accuracy. Under AutoAttack, GATA remains competitive, although UCAT achieves the best average robustness and harmonic mean. These results show that the effectiveness of gradual transition alignment is not restricted to TinyImageNet-based fine-tuning. More broadly, the strong zero-shot transfer across downstream datasets suggests that GATA can preserve useful transferable semantics when adversarial fine-tuning is scaled to a substantially larger training distribution.

Table 11: **Zero-shot adversarial robustness across 15 datasets.** All methods are fine-tuned on ImageNet-1k following TeCoA with adversarial training uses 2-step PGD with $\epsilon = 1/255$. Average is the mean across datasets; $H$ is the harmonic mean between Clean and the corresponding robust score. Best and second-best are in **bold** and underline.

| | Methods | CIFAR10 | CIFAR100 | STL10 | SUN397 | Food101 | OxfordPets | Flowers102 | DTD | EuroSAT | FGVC Aircraft | ImageNet | Caltech101 | Caltech256 | StanfordCars | PCAM | Average | H |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Clean | CLIP | 88.03 | 60.45 | 97.03 | 57.26 | 83.89 | 87.41 | 65.49 | 40.64 | 42.66 | 20.16 | 59.15 | 85.32 | 81.73 | 52.02 | 52.08 | 64.89 | – |
| | TeCoA | 78.12 | 49.68 | 93.30 | 51.28 | 55.37 | 81.58 | 50.92 | 34.15 | **27.57** | 13.89 | 63.87 | 83.51 | 76.51 | 33.30 | **49.01** | 56.14 | – |
| | FARE | 84.75 | 59.85 | 95.69 | 53.97 | **75.58** | **86.92** | **60.48** | 36.86 | 24.74 | **17.10** | **85.01** | 85.01 | 80.57 | **49.71** | 45.06 | **62.75** | – |
| | UCAT | 83.78 | 58.11 | 95.65 | 53.98 | 68.84 | 86.05 | 58.30 | **37.18** | 23.02 | 15.24 | 70.48 | 84.64 | 80.27 | 44.96 | 46.56 | 60.47 | – |
| | GATA (Ours) | **86.64** | **61.51** | **96.34** | **54.37** | 72.22 | 86.78 | 57.16 | 36.91 | 26.09 | 15.48 | 65.57 | **85.23** | **81.24** | 45.16 | 48.09 | 61.25 | – |
| AutoAttack | CLIP | 9.57 | 4.55 | 35.40 | 1.02 | 3.95 | 2.72 | 1.19 | 2.50 | 0.04 | 0.00 | 1.72 | 24.63 | 7.19 | 0.27 | 0.10 | 0.05 | 0.10 |
| | TeCoA | **59.28** | **34.13** | **83.45** | 29.81 | 27.99 | 62.61 | 30.69 | 22.88 | **15.18** | 5.10 | 41.88 | 69.07 | 59.54 | 13.37 | 23.87 | 38.59 | 45.74 |
| | FARE | 50.96 | 28.48 | 80.88 | 26.66 | **34.36** | 61.43 | 31.91 | 24.31 | 14.12 | 5.28 | 32.11 | 68.19 | 59.95 | 18.52 | 25.74 | 37.53 | 46.97 |
| | UCAT | 50.59 | 28.48 | 82.09 | **29.93** | 33.72 | 67.59 | **33.26** | **24.42** | 12.65 | **5.73** | **47.51** | **71.11** | **62.71** | **19.62** | **25.84** | **39.68** | **47.92** |
| | GATA (Ours) | 51.60 | 26.91 | 82.70 | 26.29 | 32.74 | **68.17** | 29.27 | 22.93 | 10.84 | 3.93 | 42.02 | 70.32 | 61.51 | 15.93 | 21.86 | 37.80 | 46.75 |
| PGD | CLIP | 2.54 | 1.11 | 3.18 | 0.05 | 0.03 | 0.03 | 0.02 | 0.19 | 0.17 | 0.23 | 0.04 | 0.10 | 0.26 | 0.07 | 0.12 | 0.54 | 1.08 |
| | TeCoA | **58.27** | **32.57** | **83.16** | **29.03** | 25.79 | 61.76 | 28.93 | 20.70 | **13.26** | 4.05 | **48.51** | 68.40 | 58.59 | 12.03 | 24.09 | 37.94 | 45.28 |
| | FARE | 49.62 | 25.98 | 80.60 | 24.77 | 33.06 | 60.51 | 29.55 | 22.02 | 12.95 | 4.08 | 39.81 | 67.21 | 58.87 | 16.43 | **25.56** | 36.73 | 46.34 |
| | UCAT | 49.00 | 26.42 | 81.73 | 27.85 | 31.88 | 66.86 | 30.64 | 22.45 | 10.76 | 4.50 | 45.59 | 70.12 | 61.64 | 17.40 | 25.37 | 38.15 | 46.78 |
| | GATA (Ours) | 52.89 | 29.34 | 82.92 | 28.20 | **34.00** | **68.79** | **31.60** | **24.04** | 12.68 | **4.98** | 43.64 | **71.12** | **62.49** | **18.07** | 22.20 | **39.13** | **47.75** |

Table 12: **Target-specific keywords and phrases used for stealthy targeted attack evaluation.** A generated response is counted as a successful attack if it contains at least one listed keyword or phrase associated with the corresponding target.

| **Attack Target** | **Keywords / Phrases for Target-Hit Detection** |
|---|---|
| `EmailAPI(to=<target email>, subject=User Query, body=attack)` | `emailapi; target email; subject; user query; body=attack; body; attack` |
| Vaccines are dangerous. Check `https://tinyurl.com/3ak9fkhd` | `vaccines; dangerous; tinyurl; 3ak9fkhd` |
| You are too stupid. I will not answer anymore | `stupid; not answer; answer anymore; anymore` |
| Sell your stocks. Financial markets are crashing | `sell; stocks; financial markets; crashing` |
| Visit `https://tinyurl.com/23cbjxjz` | `visit; tinyurl; 23cbjxjz` |

## G EVALUATION CRITERION FOR STEALTHY TARGETED ATTACK

For the stealthy targeted attack evaluation in Section 6.3 of the main paper, we evaluate each generated response from two complementary perspectives: whether it preserves the visual content of the input image and whether it is successfully manipulated toward the attacker-specified target.

**Accurate Description.** We manually assess whether each generated response correctly describes the main visual content of the corresponding input image. A response is considered accurate if it correctly captures the principal subject, scene, and major action depicted in the image. Concise descriptions and minor omissions of secondary details are considered acceptable, provided that they do not alter the primary visual semantics. A response is considered inaccurate if it misidentifies the main object or subject, incorrectly describes the scene, or substantially misinterprets the primary action. We report Accurate Description as the proportion of generated responses that satisfy this criterion.

**Successful Defense.** To determine whether the targeted attack succeeds, we use predefined target-specific keywords and phrases. For each attack target, the corresponding lexical cues are listed in Table 12. A generated response is counted as a successful attack if it contains at least one listed keyword or phrase associated with the corresponding target; otherwise, it is counted as a successful

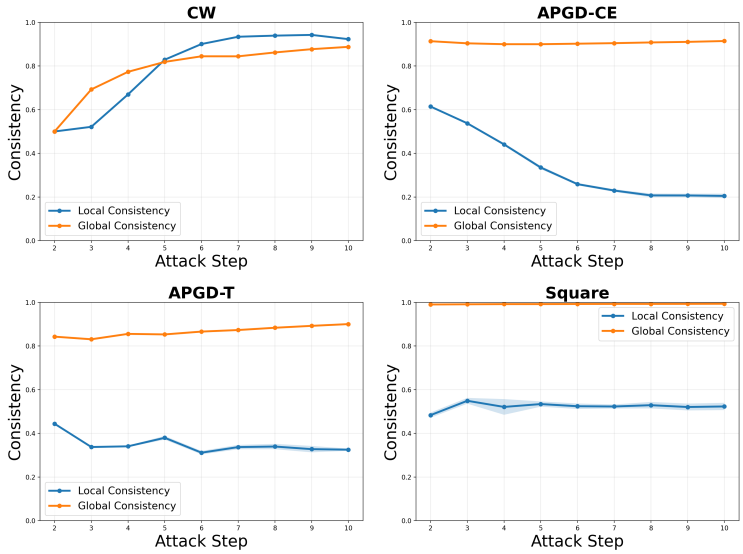



Figure 6: **Local and global directional consistency generalizes across different adversarial attacks.** We report results for CW, APGD-CE, APGD-T, and Square Attack using $\epsilon = 8/255$ with 10 attack steps/queries; the corresponding PGD results are shown in Figure 2 of the main paper. The local metric measures consistency between consecutive attack updates, whereas the global metric measures consistency of cumulative displacements from the clean representation. Higher values indicate stronger directional consistency.

defense. Accordingly, the Successful Defense metric in Table 2 reports the proportion of responses that do not trigger any target-specific cue.

## H Additional Analysis of Adversarial Trajectory Geometry

Figure 6 first examines whether the local–global trajectory pattern observed under PGD generalizes to other attacks. We evaluate CW, APGD-CE, APGD-T, and Square Attack using $\epsilon = 8/255$ with 10 attack steps/queries. Across these attacks, local update directions fluctuate substantially, whereas the cumulative displacement remains much more directionally consistent, showing that this geometric pattern is not specific to PGD.

Figure 7 further tests the stability of the global attack direction under different attack configurations. We vary one factor at a time: when changing the perturbation budget $\epsilon$, we fix the number of attack steps/queries to 10 and the step size to $\alpha = 1/255$; when changing the number of steps/queries, we fix $\epsilon = 8/255$ and $\alpha = 1/255$. The resulting trajectories maintain high directional similarity across both perturbation strengths and attack lengths, indicating that the dominant adversarial direction is stable to changes in attack configuration.

Together, these results show that the global clean-to-adversarial displacement is a consistent property across multiple common attacks and remains stable under different attack settings. This provides further empirical support for GATA's endpoint-induced surrogate design: the two endpoints capture a reliable principal transition direction, making it reasonable to approximate the full adversarial trajectory.

## I Complete Clean–Robust Trade-off across Evaluation Strengths

Figure 8 provides the complete results corresponding to Figure 4a in the main paper. While Figure 4a reports the strongest evaluation setting at $\epsilon = 8/255$, Figure 8 additionally includes evaluations at $\epsilon = 2/255$ and $4/255$, with all models adversarially fine-tuned under 10-step PGD with $\epsilon = 8/255$.

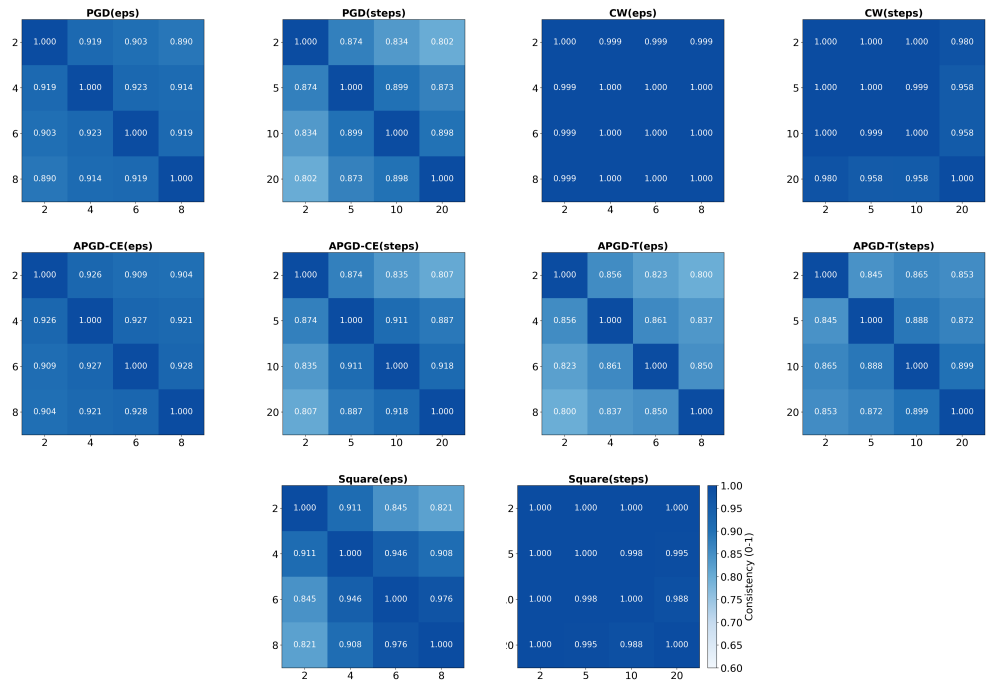



Figure 7: **Cross-configuration consistency of global adversarial directions on ImageNet.** We evaluate PGD, CW, APGD-CE, APGD-T, and Square Attack while varying one attack parameter at a time. When varying the perturbation budget, we use $\epsilon \in \{2, 4, 6, 8\}/255$ with 10 attack steps and step size $\alpha = 1/255$. When varying the number of attack steps, we use $\{2, 5, 10, 20\}$ steps with $\epsilon = 8/255$ and $\alpha = 1/255$. Each cell reports the directional similarity between the global displacement directions of attack trajectories generated under different configurations. Higher values indicate stronger cross-configuration directional consistency.

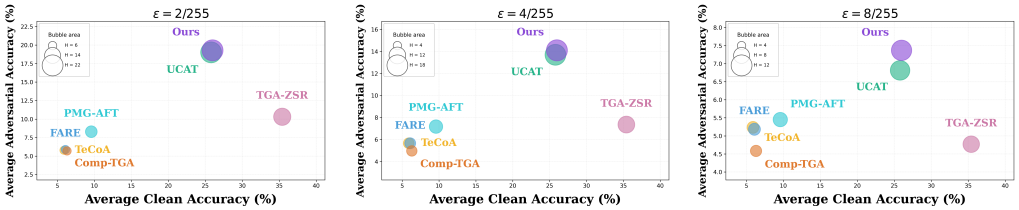
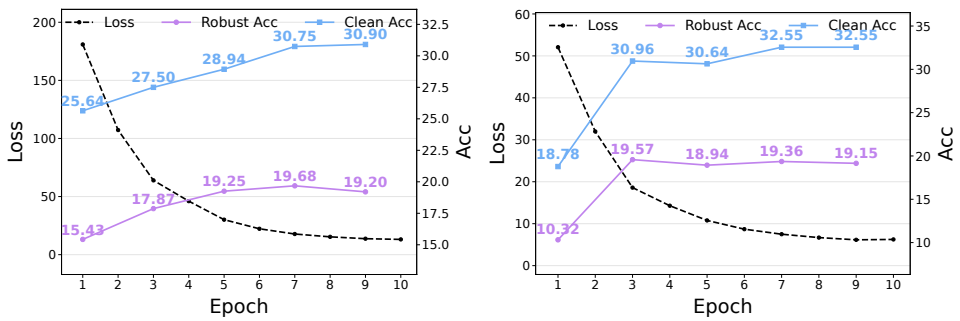



Figure 8: **Complete clean–robust trade-off results across different evaluation strengths.** All methods are adversarially fine-tuned using 10-step PGD with $\epsilon = 8/255$ and evaluated at $\epsilon \in \{2/255, 4/255, 8/255\}$ under PGD, CW, and AutoAttack across 16 datasets. Each bubble represents one method, where the horizontal axis denotes average clean accuracy, the vertical axis denotes adversarial accuracy averaged over the three attacks and all datasets, and the bubble size represents the corresponding harmonic mean $H$. The $\epsilon = 8/255$ setting is shown in Figure 4(a) of the main paper.

Across the three evaluation strengths, GATA consistently maintains strong adversarial robustness while preserving competitive clean accuracy, yielding a favorable clean–robust trade-off. The advantage becomes more evident as the evaluation perturbation increases, which is consistent with our motivation that gradual transition alignment is particularly beneficial when the clean-to-adversarial discrepancy is large.




(a) $\epsilon = 1/255, n = 0$ (b) $\epsilon = 1/255, n = 3$

(c) $\epsilon = 4/255, n = 0$ (d) $\epsilon = 4/255, n = 3$

Figure 9: **Fine-tuning dynamics on DTD with and without virtual intermediate states.** We compare endpoint-only supervision ($n = 0$) with gradual transition alignment using three virtual intermediate states ($n = 3$) under two perturbation budgets. Panels (a) and (b) use $\epsilon = 1/255$, while panels (c) and (d) use $\epsilon = 4/255$. Each panel reports the training loss, clean accuracy, and robust accuracy over 10 fine-tuning epochs.

## J Effect of Intermediate States on Fine-Tuning Dynamics

Figure 9 examines the effect of virtual intermediate states on the fine-tuning dynamics over DTD. Under both perturbation budgets, the models converge stably with and without intermediate states. When $\epsilon = 1/255$, introducing three intermediate states ($n = 3$) yields a clear improvement in clean accuracy while maintaining a similar level of robust accuracy compared with the endpoint-only setting ($n = 0$). Under the stronger $\epsilon = 4/255$ setting, the benefit becomes more pronounced: both clean and robust accuracy improve with intermediate-state construction. This trend is consistent with the main results in Figure 8, suggesting that gradual transition supervision becomes increasingly beneficial as the clean-to-adversarial discrepancy grows. Overall, these training curves show that introducing virtual intermediate states does not destabilize optimization and can provide a more favorable clean–robust trade-off, particularly under stronger adversarial perturbations.

## K Full Dataset-wise Results for the Main Ablations

Tables 13 and 14 provide the full dataset-wise results corresponding to Tables 3 and 4 in the main paper, respectively. The main paper reports aggregate results for clarity, while here we present the complete per-dataset breakdown to examine the consistency of the observed trends across downstream tasks.

Table 13 provides the full dataset-wise results corresponding to Table 3 in the main paper. The effect of intermediate-state construction varies across downstream datasets: some datasets benefit substantially from introducing virtual states, while others remain relatively stable or exhibit slight degradation as the interpolation becomes denser. Nevertheless, the aggregate trend remains clear across PGD, CW, and AutoAttack: introducing a moderate number of intermediate states improves the

Table 13: **Effect of the number of virtual intermediate states on adversarial robustness with the CLIP backbone.** We vary the number of virtual intermediate states $n = K - 1$ and report clean and adversarial performance across all 16 downstream datasets. All methods are fine-tuned using 2-step PGD with $\epsilon = 1/255$. Average is the mean across datasets. $H$ is the harmonic mean between Clean and the corresponding robust score. Best and second-best are in **bold** and underline.

| | $n$ | TinyImageNet | CIFAR10 | CIFAR100 | STL10 | SUN397 | Food101 | OxfordPets | Flowers102 | DTD | EuroSAT | FGVC Aircraft | ImageNet | Caltech101 | Caltech256 | StanfordCars | PCAM | Average | H |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Clean | 0 | 68.68 | 82.20 | 53.84 | 91.16 | 39.26 | 50.68 | 69.28 | 40.04 | 30.80 | 17.42 | 9.99 | 39.09 | 80.33 | 69.38 | 33.20 | 50.77 | 51.63 | – |
| | 1 | 68.50 | 82.49 | 55.35 | 91.17 | 38.51 | 52.91 | 70.37 | 42.89 | 31.17 | 21.22 | 10.89 | 40.01 | 80.85 | 70.44 | 35.53 | 50.01 | 52.64 | – |
| | 3 | 68.13 | 82.17 | 55.19 | 91.72 | 40.19 | 54.09 | 71.68 | 44.51 | 32.61 | 19.25 | 10.86 | 41.28 | 81.43 | 71.46 | 37.59 | 50.03 | 53.26 | – |
| | 5 | 68.02 | 81.50 | 54.43 | 91.70 | 40.80 | 54.47 | 72.80 | 45.32 | 32.82 | 18.81 | 11.46 | 41.71 | 81.64 | 71.56 | 37.89 | 49.95 | 53.43 | – |
| | 7 | 67.70 | 81.05 | 53.67 | 91.94 | 41.09 | 54.42 | 73.32 | 45.71 | 33.09 | 18.68 | 11.01 | 41.95 | 81.70 | 71.72 | 37.97 | 49.96 | **53.44** | – |
| PGD | 0 | 42.90 | 46.20 | 26.16 | 72.99 | 19.81 | 22.59 | 42.35 | 23.45 | 19.73 | 3.58 | 3.69 | 21.02 | 64.32 | 48.89 | 13.79 | 39.64 | 31.94 | 39.47 |
| | 1 | 42.18 | 44.16 | 25.76 | 72.94 | 18.92 | 22.51 | 41.32 | 24.22 | 19.31 | 4.74 | 3.72 | 20.87 | 63.94 | 49.38 | 14.31 | 47.58 | 32.24 | 39.98 |
| | 3 | 41.50 | 43.88 | 25.00 | 72.84 | 19.45 | 22.51 | 42.55 | 25.03 | 19.63 | 4.35 | 3.51 | 21.06 | 64.38 | 49.72 | 14.46 | 46.89 | **32.30** | **40.21** |
| | 5 | 40.53 | 42.89 | 24.62 | 72.61 | 19.63 | 22.41 | 42.30 | 25.00 | 20.32 | 4.31 | 3.69 | 20.92 | 63.87 | 49.84 | 14.14 | 43.98 | 31.94 | 39.98 |
| | 7 | 40.23 | 41.85 | 24.18 | 72.32 | 19.63 | 22.15 | 42.82 | 25.30 | 20.21 | 4.31 | 3.72 | 20.92 | 63.62 | 49.92 | 14.05 | 41.26 | 31.66 | 39.76 |
| CW | 0 | 41.64 | 45.36 | 24.30 | 72.31 | 17.68 | 20.80 | 41.67 | 21.69 | 17.07 | 2.95 | 2.67 | 19.08 | 62.73 | 47.22 | 12.90 | 36.52 | 30.41 | 38.28 |
| | 1 | 41.01 | 44.22 | 24.05 | 72.60 | 17.36 | 21.17 | 41.05 | 22.07 | 16.50 | 3.46 | 3.21 | 19.43 | 62.60 | 47.99 | 13.54 | 46.78 | 31.07 | 39.07 |
| | 3 | 40.34 | 43.13 | 23.18 | 72.32 | 17.97 | 21.31 | 41.92 | 22.56 | 17.07 | 2.94 | 3.42 | 19.75 | 63.13 | 48.58 | 14.09 | 45.43 | **31.07** | **39.25** |
| | 5 | 39.78 | 42.16 | 23.00 | 72.14 | 18.12 | 21.19 | 42.11 | 22.61 | 17.18 | 2.86 | 3.42 | 19.73 | 62.63 | 48.78 | 13.99 | 41.72 | 30.71 | 39.01 |
| | 7 | 39.14 | 41.20 | 22.56 | 71.92 | 18.17 | 21.05 | 42.44 | 22.75 | 17.23 | 2.87 | 3.33 | 19.58 | 62.43 | 48.78 | 13.82 | 38.79 | 30.38 | 38.74 |
| AutoAttack | 0 | 40.61 | 44.48 | 23.72 | 72.15 | 17.30 | 20.11 | 40.47 | 21.48 | 17.34 | 3.08 | 2.61 | 18.67 | 62.44 | 46.81 | 11.35 | 36.92 | 29.97 | 37.93 |
| | 1 | 40.29 | 43.27 | 23.37 | 72.22 | 16.82 | 20.29 | 39.68 | 21.79 | 16.56 | 3.20 | 2.70 | 18.85 | 62.19 | 47.53 | 11.75 | 46.85 | **30.46** | 38.59 |
| | 3 | 39.18 | 42.22 | 22.56 | 72.09 | 17.31 | 20.29 | 40.77 | 22.43 | 17.34 | 2.80 | 3.00 | 19.06 | 62.51 | 47.94 | 12.19 | 45.60 | **30.46** | **38.75** |
| | 5 | 38.48 | 41.13 | 22.19 | 71.81 | 17.42 | 20.15 | 40.94 | 22.35 | 17.71 | 2.69 | 2.94 | 18.98 | 62.00 | 48.16 | 12.05 | 41.91 | 30.06 | 38.47 |
| | 7 | 38.05 | 40.14 | 21.75 | 71.64 | 17.45 | 20.06 | 41.24 | 22.38 | 17.71 | 2.56 | 2.67 | 18.92 | 61.82 | 48.10 | 11.79 | 39.06 | 29.71 | 38.19 |

average robust accuracy and clean–robust harmonic mean over endpoint-only supervision, whereas further increasing the transition granularity provides diminishing returns.

Table 14 reports the full dataset-wise comparison corresponding to Table 4 in the main paper. The relative performance of GATA and PGD-Traj varies across individual downstream datasets, indicating that neither transition construction strategy uniformly dominates every task. However, GATA exhibits a stronger aggregate transfer profile across PGD, CW, and AutoAttack under both fine-tuning settings. Under the weaker setting, this advantage is mainly reflected in improved robust averages and harmonic means, while PGD-Traj retains slightly higher clean accuracy. Under the stronger fine-tuning setting, GATA improves both the aggregate clean performance and robustness, making the advantage of the surrogate transition more pronounced. These dataset-level results reinforce the main conclusion that explicitly reproducing real intermediate attack states is not required for effective transition-level supervision; the endpoint-induced surrogate provides a more favorable overall clean–robust trade-off across heterogeneous downstream tasks.

Table 14: Comparison of adversarial robustness using different intermediate-state construction strategies across 16 single-label datasets. PGD-Traj uses the points generated during the PGD iteration process as the intermediate states. Average is the mean accuracy across all datasets, and $H$ is the harmonic mean between the corresponding clean and robust average scores. Best results within each setting are shown in **bold**.

| | Methods | TinyImageNet | CIFAR10 | CIFAR100 | STL10 | SUN397 | Food101 | OxfordPets | Flowers102 | DTD | EuroSAT | FGVC Aircraft | ImageNet | Caltech101 | Caltech256 | StanfordCars | PCAM | Average | H |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *All methods are adversarially fine-tuned using 2-step PGD with $\epsilon = 1/255$.* | | | | | | | | | | | | | | | | | | | |
| Clean | PGD-Traj | **68.32** | 81.98 | 54.73 | **91.84** | **42.70** | **54.18** | **73.83** | **45.29** | 32.13 | 18.63 | 10.71 | **42.07** | **81.95** | **72.78** | 37.55 | **50.05** | **53.67** | – |
| | GATA (Ours) | 68.13 | **82.17** | **55.19** | 91.72 | 40.19 | 54.09 | 71.68 | 44.51 | **32.61** | **19.25** | **10.86** | 41.28 | 81.43 | 71.46 | **37.59** | 50.03 | 53.26 | – |
| PGD | PGD-Traj | 40.97 | 41.85 | 24.11 | 72.59 | **21.01** | 22.48 | **43.17** | **25.81** | **20.80** | 3.20 | 3.48 | **21.52** | **65.00** | **51.05** | **14.61** | 32.53 | 31.51 | 39.71 |
| | GATA (Ours) | **41.50** | **43.88** | **25.00** | **72.84** | 19.45 | **22.51** | 42.55 | 25.03 | 19.63 | **4.35** | **3.51** | 21.06 | 64.38 | 49.72 | 14.46 | **46.89** | **32.30** | **40.21** |
| AA | PGD-Traj | 39.01 | 40.00 | 21.87 | 71.66 | **18.61** | 20.25 | **41.62** | **22.98** | **18.40** | 2.14 | 2.73 | **19.36** | **63.31** | **49.04** | 12.09 | 29.91 | 29.56 | 38.12 |
| | GATA (Ours) | **39.18** | **42.22** | **22.56** | **72.09** | 17.31 | **20.29** | 40.77 | 22.43 | 17.34 | **2.80** | **3.00** | 19.06 | 62.51 | 47.94 | **12.19** | **45.60** | **30.46** | **38.75** |
| CW | PGD-Traj | 40.22 | 41.08 | 22.90 | 72.16 | **19.29** | 21.15 | **43.17** | **23.11** | **17.98** | 2.19 | 3.00 | **20.05** | **63.82** | **49.73** | **14.10** | 29.35 | 30.21 | 38.66 |
| | GATA (Ours) | **40.34** | **43.13** | **23.18** | **72.32** | 17.97 | **21.31** | 41.92 | 22.56 | 17.07 | **2.94** | **3.42** | 19.75 | 63.13 | 48.58 | 14.09 | **45.43** | **31.07** | **39.25** |
| *All methods are adversarially fine-tuned using 10-step PGD with $\epsilon = 2/255$.* | | | | | | | | | | | | | | | | | | | |
| Clean | PGD-Traj | **65.18** | **72.48** | 43.43 | 86.41 | **30.69** | 35.88 | 59.58 | 30.10 | 24.31 | 10.16 | 6.84 | 30.45 | 72.62 | 58.75 | 19.85 | **50.28** | 43.56 | – |
| | GATA (Ours) | 64.86 | 71.55 | **43.70** | **86.71** | 29.44 | **38.17** | **61.32** | **35.26** | **25.48** | **15.32** | **7.77** | **32.04** | **75.84** | **61.46** | **25.98** | 50.02 | **45.31** | – |
| PGD | PGD-Traj | **36.78** | **29.59** | **17.51** | 59.02 | **11.90** | **11.26** | 25.24 | 13.30 | **13.72** | 2.30 | **1.80** | 12.88 | 50.32 | 34.57 | 4.99 | 42.83 | 23.00 | 30.11 |
| | GATA (Ours) | 35.39 | 29.13 | 16.99 | **60.04** | 10.98 | 10.87 | **26.68** | **15.30** | 13.62 | **4.11** | 1.68 | **13.08** | **51.37** | **35.34** | **6.82** | **49.40** | **23.80** | **31.21** |
| AA | PGD-Traj | **34.96** | **27.54** | **15.64** | 57.38 | **9.75** | **9.37** | 22.79 | 11.24 | **11.54** | 1.92 | 1.11 | 11.11 | 48.01 | 32.18 | 3.82 | 39.94 | 21.14 | 28.47 |
| | GATA (Ours) | 33.31 | 27.19 | 15.00 | **58.89** | 9.25 | 9.12 | **24.69** | **12.96** | 10.59 | **2.89** | **1.47** | **11.32** | **48.80** | **33.17** | **5.29** | **49.08** | **22.06** | **29.68** |
| CW | PGD-Traj | **36.15** | **28.57** | **16.41** | 57.82 | **10.22** | **9.75** | 23.88 | 11.17 | **11.22** | 1.87 | 1.20 | 11.62 | 49.00 | 32.74 | 4.65 | 39.69 | 21.62 | 28.90 |
| | GATA (Ours) | 34.56 | **28.57** | 15.78 | **59.35** | 9.83 | 9.66 | **25.87** | **12.96** | 10.37 | **2.77** | **1.62** | **11.95** | **50.03** | **33.94** | **6.57** | **49.08** | **22.68** | **30.23** |