`Robust SuperAlignment`: Weak-to-Strong Robustness Generalization for Vision-Language Models

Junhao Dong[1,2*], Cong Zhang[3†], Xinghua Qu[4], Zejun Ma[3],
Piotr Koniusz[5,6,7], and Yew-Soon Ong[1,2†]
[1]Nanyang Technological University, [2]CFAR, IHPC, A*STAR, [3]TikTok, Singapore,
[4]Bytedance, [5]Data61♥CSIRO, [6]University of New South Wales, [7]Australian National University

Abstract

Numerous well-established studies have demonstrated the superhuman capabilities of modern Vision-Language Models (VLMs) across a wide range of tasks. However, growing is the doubt about the continuing availability of reliable high-quality labeling (supervision) from human annotators, leading to stagnation of the model's performance. To address this challenge, "superalignment" employs the so-called weak-to-strong generalization paradigm, where the supervision from a *weak* model can provide generalizable knowledge for a *strong* model. While effective in aligning knowledge for clean samples between the strong and weak models, the standard weak-to-strong approach typically fails to capture adversarial robustness, exposing strong VLMs to adversarial attacks. This inability to transfer adversarial robustness is because adversarial samples are normally missing in the superalignment stage. To this end, we are the first to propose the weak-to-strong (adversarial) robustness generalization method to elicit zero-shot robustness in large-scale models by an **unsupervised** scheme, mitigating the unreliable information source for alignment from two perspectives: *alignment re-weighting* and *source guidance refinement*. We analyze settings under which robustness generalization is possible. Extensive experiments across various vision-language benchmarks validate the effectiveness of our method in numerous scenarios, demonstrating its plug-and-play applicability to large-scale VLMs.

1 Introduction

Vision-Language Models (VLMs) enjoy remarkable prediction capabilities, frequently surpassing human performance on diverse multimodal tasks, ranging from zero-shot recognition to multimodal reasoning [60, 42, 49] to task unlearning [24]. Despite these outstanding achievements, they suffer from an emerging issue: as models become highly capable, human annotators fail to reliably guide and/or evaluate their outputs, especially given the complexity and ambiguity inherent in vision-language paired data due to inherent limitations in cognitive capacity and scalability [7, 4]. This fundamental limitation has prompted the super-intelligence alignment (*superalignment*) [41], a conceptualized framework to align (future) superhuman models with human values and goals based on limited human supervision. One seminal approach toward superalignment is *weak-to-strong generalization* [5], which explores supervision from a weak (lightweight) *teacher* model to guide a strong (large-scale) *student* model, enabling the student to generalize beyond the original limitations of its supervision source (reference model) on unforeseen data.

However, while existing weak-to-strong generalization methods focus on distilling knowledge from a teacher model, we demonstrate that robustness, a distinct form of knowledge, cannot be transferred

[*]The research work was done during Junhao's internship at TikTok Singapore.
[†]Corresponding author.

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

into a VLM even when guided by a robust teacher[1]. This limitation inherently exacerbates the vulnerability to adversarial examples—original images with added imperceptible artifacts leading to erroneous predictions with high confidence [65]. In addition to single-modal backbones [21, 20, 23, 15, 22, 14], these vulnerabilities are especially critical in VLMs, where the correctness of language reasoning heavily relies on visual inputs: poisoned images may compound errors and propagate catastrophic failures through the model's output, fundamentally compromising its reliability for real-world deployment [72, 51, 46, 47].

Through systematic analysis, we are the first to identify the root cause of robustness transfer failure presenting as a mismatch in learning objectives arising from the data employed in weak-to-strong generalization models. Our analysis in Section 3.2 reveals that integrating adversarial examples—rather than their clean counterparts—into learning can induce robustness in the strong student model, even without explicit supervision from an adversarially robust teacher. Motivated by this insight, we propose the first adversarially robust weak-to-strong generalization framework, termed **Adv-W2S**, designed to elicit robust knowledge from the strong student model in an unsupervised scheme. Specifically, our method mitigates unreliable supervision signals from the weak teacher model through two complementary mechanisms: *alignment re-weighting* and *source guidance refinement*. To quantify the reliability of guidance of the weak teacher, we introduce an instance-wise re-weighting mechanism based on prediction entropy, adaptively modulating the emphasis placed on robust knowledge alignment between the teacher and student VLMs. Additionally, in order to further reduce reliance on potentially erroneous guidance signals, we design an adaptive refinement scheme for prediction reference, generating benign inputs that can be considered as improved guidance for robust weak-to-strong generalization. Our theoretical analysis establishes two core claims: (i) low prediction entropy enforced on the student promotes larger classification margins and robustness, and (ii) alignment toward refined source guidance strictly improves robustness generalization during superalignment.

Extensive experiments conducted across diverse network architectures and evaluation scenarios demonstrate that our proposed Adv-W2S framework achieves superior zero-shot classification performance in both clean and robust accuracy compared to state-of-the-art adversarial fine-tuning methods. Moreover, we show that the robustness induced by Adv-W2S can effectively transfer to a broad array of downstream vision-language tasks—including image captioning, visual question answering, hallucination mitigation, and chain-of-thought reasoning—via a **plug-and-play** replacement of the original vision encoder with our robustly aligned encoder. Our empirical analyses further reveal that the in-distribution robustness initially obtained via unsupervised alignment can be further improved through supervised fine-tuning. To our best knowledge, this work represents the first systematic exploration of robust superalignment, providing a promising pathway toward building robust and aligned foundation VLMs for artificial general intelligence applications.

Our core contributions are summarized as follows:

1. We reveal that standard weak-to-strong generalization schemes fail to transfer VLM robustness. Through systematic analyses, we identify—for the first time—that the root cause of this failure is the absence of adversarial examples in alignment objectives of weak-to-strong generalization.
2. In contrast to adversarial fine-tuning from scratch, we propose **Adv-W2S**, the first adversarially robust weak-to-strong generalization framework to elicit robust knowledge from a weak student VLM by an unsupervised scheme. We also investigate unreliable source guidance mitigation from two complementary perspectives: alignment re-weighting and source guidance refinement. Theoretical analyses characterize the conditions under which robustness generalization is possible.
3. We conduct extensive experiments across 20 datasets spanning diverse vision-language tasks (including visual question answering and captioning) across various scenarios, demonstrating that our Adv-W2S consistently outperforms state-of-the-art adversarial fine-tuning approaches in terms of natural performance and zero-shot robustness, while supporting plug-and-play integration.

2 Related Works

Foundation VLMs. Vision-language pre-training learns joint visual-textual representations to improve downstream task performance [73]. CLIP [60] pioneered large-scale vision-language contrastive learning, enabling zero-shot transfer to a wide range of vision tasks. Subsequent foundation

[1]Empirical evidence for our motivation is provided in Section 3.2.

VLMs, *e.g.*, BLIP [43], OpenFlamingo [2], and LLaVA [49, 48], have demonstrated strong multimodal understanding and generalization. Our study concentrates on improving the robustness of the widely adopted CLIP model, generalizing robustness to other foundation VLMs via a plug-and-play replacement of their vision encoders with our robust counterpart. To mitigate computational costs and potential overfitting associated with full fine-tuning, we explore Parameter-Efficient Fine-Tuning (PEFT) [36, 35, 56, 76, 74] within our robust weak-to-strong generalization framework.

Weak-to-strong generalization for superalignment. Inspired by the challenge of aligning superhuman models via weaker supervision (superalignment), Burns *et al.* [5] first explored fine-tuning large-scale models using weak-model supervision. This paradigm, termed weak-to-strong generalization, improves the generalizability of the strong student model via a weak teacher model, distinct from standard knowledge distillation [34]. Lang *et al.* [40] established the error bound for weak supervision, highlighting the significance of pseudolabel correction. Our study explores prediction-based pseudolabel refinement by generating more benign inputs for improved supervision via an inverse procedure of adversarial generation. Subsequent works studied the inherent reliability of weak teacher supervision [29, 30]. However, prior works focus on single-modal vanilla knowledge transfer for specific tasks. In contrast, we address an underexplored problem: robust weak-to-strong generalization for VLMs, generalizing natural and robust knowledge across diverse multimodal tasks.

Adversarial robustness of VLMs. The increasing security risks posed by adversarial examples [37, 70] have spurred extensive research on defense schemes for VLMs [63, 75, 3, 17, 18, 16, 19]. Adversarial fine-tuning, a leading defense paradigm, augments training data with adversaries based on naturally pre-trained VLMs, *e.g.*, CLIP [60]. Mao *et al.* [55] first introduced adversarial fine-tuning within a contrastive learning framework to align adversarially perturbed image embeddings with their text counterparts. Schlarmann *et al.* [63] developed an unsupervised adversarial fine-tuning strategy by preserving the features of the original CLIP model. Unlike prior adversarial fine-tuning approaches that exclusively rely on training data, we focus on a weak-to-strong generalization framework guided by a weak VLM to elicit the strong robustness generalization capability from a strong VLM across diverse downstream tasks, offering a promising solution toward the superalignment challenge for foundation VLMs.

3 Robust Weak-to-strong Generalization for Superalignment

Below, we propose **Adv-W2S**, the first robust weak-to-strong generalization framework to elicit zero-shot robustness from large VLMs across a variety of vision-language tasks without extra supervision.

3.1 Revisiting Adversarial Fine-tuning and Weak-to-strong Generalization

Adversarial Fine-tuning. CLIP [60] consists of an image encoder $f_{\boldsymbol{\theta}_{\mathrm{I}}}: \mathcal{X} \to \mathbb{R}^d$ and a text encoder $f_{\boldsymbol{\theta}_{\mathrm{T}}}: \mathcal{T} \to \mathbb{R}^d$, parameterized by $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\mathrm{I}}, \boldsymbol{\theta}_{\mathrm{T}})$. These encoders map image-text pairs $(\mathbf{x}, \mathbf{t})$ into d-dimensional features. Classification is performed via computing the probability that input $\mathbf{x}$ belongs to class $c \in \{1, \ldots, C\}$ via softmax over the image-text feature cosine similarity:

$$p_c(\mathbf{x}) = \frac{\exp\big(\cos(f_{\boldsymbol{\theta}_{\mathrm{I}}}(\mathbf{x}), f_{\boldsymbol{\theta}_{\mathrm{T}}}(\mathbf{t}_c))\big)}{\sum_{c'=1}^{C} \exp\big(\cos(f_{\boldsymbol{\theta}_{\mathrm{I}}}(\mathbf{x}), f_{\boldsymbol{\theta}_{\mathrm{T}}}(\mathbf{t}_{c'}))\big)}, \tag{1}$$

where $\exp(\cdot)$ and $\cos(\cdot)$ denote the exponential and cosine similarity functions. Each text prompt is tokenized by $g(\cdot)$ and embedded by $f_{\boldsymbol{\theta}_{\mathrm{T}}}(\cdot)$, denoted as $\mathbf{t}_{c'} = g($"$\texttt{[Context][CLASS}_{c'}\texttt{]}$"$)$. A typical template is "$\texttt{This is a photo of[CLASS}_{c'}\texttt{]}$". We define $\mathbf{p}(\mathbf{x}) = [p_1(\mathbf{x}), \ldots, p_C(\mathbf{x})]^\top \in [0,1]^C$ as the prediction for input $\mathbf{x}$ across C categories. Given image-text pairs $\mathcal{D}$, standard adversarial fine-tuning (TeCoA) [55] is framed as a minimax optimization to improve CLIP robustness:

$$\min_{\boldsymbol{\theta}_{\mathrm{I}}} \mathbb{E}_{(\mathbf{x},c)\sim\mathcal{D}} \left[\max_{\|\boldsymbol{\delta}\|_\infty \le \epsilon} \mathcal{L}_{\mathrm{CE}}\big(\mathbf{p}(\mathbf{x}+\boldsymbol{\delta}), \mathbf{y}(c)\big) \right], \tag{2}$$

where $\mathbf{y}(c) = [\mathbb{1}(c=1), \ldots, \mathbb{1}(c=C)]^\top \in \{0,1\}^C$ is a one-hot label for class c, and $\mathcal{L}_{\mathrm{CE}}$ is the Cross-Entropy (CE) loss. The adversarial example $\hat{\mathbf{x}} = \mathbf{x} + \boldsymbol{\delta}$ lies within an ℓ_∞-norm hyperball of radius ϵ around $\mathbf{x}$. Further details about adversarial learning are in Appendix A. CLIP parameters are optimized by empirical risk on adversaries, generated via Projected Gradient Descent (PGD) [6].

$$\hat{\mathbf{x}}^{(i+1)} = \Pi_{\mathbb{B}(\mathbf{x},\epsilon)} \left[\hat{\mathbf{x}}^{(i)} + \alpha \cdot \mathrm{sign}\left(\nabla_{\hat{\mathbf{x}}^{(i)}} \mathcal{L}_{\mathrm{CE}}\big(\mathbf{p}(\hat{\mathbf{x}}^{(i)}), \mathbf{y}(c)\big)\right)\right], \tag{3}$$

where $\mathrm{sign}(\cdot)$ represents the sign function, and the scalar α denotes the step size. The projection operator $\Pi_{\mathbb{B}(\mathbf{x},\epsilon)}$ restricts the adversary in ℓ_∞-norm hyperball with ϵ-radius around $\mathbf{x}$. The starting point is randomly initialized $\hat{\mathbf{x}}^{(0)} \sim \mathbf{x} + 0.001 \cdot \mathcal{N}(\mathbf{0}, \mathbf{I})$ with the final adversary $\hat{\mathbf{x}} = \hat{\mathbf{x}}^{(m)}$ after m steps.

Weak-to-strong Generalization. In analogy to superalignment, Burns *et al.* [5] proposed weak-to-strong generalization via fine-tuning a strong student model using soft-label supervision from a weak teacher. We extend this idea to VLMs, denoting the weak teacher VLM as $[f_{\boldsymbol{\theta}_\mathrm{I}^{\mathcal{T}}}(\cdot), f_{\boldsymbol{\theta}_\mathrm{T}^{\mathcal{T}}}(\cdot)]$ and the strong student VLM as $[f_{\boldsymbol{\theta}_\mathrm{I}^{\mathcal{S}}}(\cdot), f_{\boldsymbol{\theta}_\mathrm{T}^{\mathcal{S}}}(\cdot)]$. Based on Eq. (1), let $\mathbf{p}_{\mathcal{T}}(\mathbf{x})$ and $\mathbf{p}_{\mathcal{S}}(\mathbf{x})$ denote the prediction of the teacher and student on input $\mathbf{x}$. The standard weak-to-strong generalization approach is formulated as an auxiliary confidence loss:

$$\mathcal{L}_{\text{AuxConf}} = (1-\beta) \cdot \mathcal{L}_{\text{CE}}\big(\mathbf{p}_{\mathcal{S}}(\mathbf{x}), \mathbf{p}_{\mathcal{T}}(\mathbf{x})\big) + \beta \cdot \mathcal{L}_{\text{CE}}\big(\mathbf{p}_{\mathcal{S}}(\mathbf{x}), \mathcal{M}(\mathbf{p}_{\mathcal{S}}(\mathbf{x}))\big), \tag{4}$$

where $\mathcal{M}(\cdot)$ produces one-hot labels based on strong model predictions, and β controls the trade-off between teacher-student prediction alignment (first term) and student self-refinement (second term).

Problem definition. Unlike standard robustness evaluations that assume in-distribution adversarial attacks [11], we study a more challenging *zero-shot* robustness scenario [55], where adversaries originate from unseen distributions during inference. Under practical defense conditions, we presume textual prompts are fixed and safeguarded, as they typically reside within multimodal systems and are thus not subjected to manipulation. Within this demanding zero-shot robustness context, we further explore how weak-to-strong generalization contributes to improved robustness elicitation across different vision-language model settings and downstream-task generalization.

3.2 Can Vanilla Weak-to-strong Generalization Elicit Robustness?

While vanilla weak-to-strong generalization improves natural performance in single modality tasks [5], its effectiveness in enhancing VLM robustness remains underexplored. We investigate whether the vanilla weak-to-strong generalization scheme (Eq. (4)) can elicit robustness from the strong student model. As shown in Table 1, standard weak-to-strong generalization primarily improves natural performance, but fails to yield robustness in large-scale student VLMs. This contrasts with vanilla knowledge distillation, where robustness has been successfully transferred from robust teachers [27]. Intriguingly, even substituting the weak guidance with an adversarially robust VLM (*e.g.*, TeCoA [55]) does not facilitate robustness elicitation. We attribute this to a mismatch in model capacity and learning objectives. Representations from a limited-capacity VLM trained on clean data poorly generalize to a high-capacity VLM tasked with handling unforeseen adversaries.

Table 1: Performance (%) of fine-tuned teacher CLIP models, and their weak-to-strong generalizations, where robustness is w.r.t. Auto-Attack [12]. $\xrightarrow{\text{Vanilla-W2S}}$ and $\xrightarrow{\text{Adversarial-W2S}}$ denote vanilla and adversarial weak-to-strong generalization.

Method	ImageNet		Avg. 13 Datasets	
	Clean	Robust	Clean	Robust
Natural Fine-Tuning	76.06	0.00	58.64	0.01
Robust Fine-Tuning	64.96	39.74	48.27	32.92
Natural FT $\xrightarrow{\text{Vanilla-W2S}}$	79.54	0.00	65.50	0.26
Robust FT $\xrightarrow{\text{Vanilla-W2S}}$	72.72	5.60	55.37	9.20
Natural FT $\xrightarrow{\text{Adversarial-W2S}}$	76.10	50.82	63.28	40.97
Robust FT $\xrightarrow{\text{Adversarial-W2S}}$	74.95	51.97	62.57	41.70

To validate our claim, we explore an adversarial variant of vanilla weak-to-strong generalization by shifting the data perspective of the strong student VLM to adversarial samples (colored in red):

$$\mathcal{L}_{\text{Adv-AuxConf}} = (1-\beta) \cdot \mathcal{L}_{\text{CE}}\big(\mathbf{p}_{\mathcal{S}}(\mathbf{x}+\boldsymbol{\delta}), \mathbf{p}_{\mathcal{T}}(\mathbf{x})\big) + \beta \cdot \mathcal{L}_{\text{CE}}\big(\mathbf{p}_{\mathcal{S}}(\mathbf{x}+\boldsymbol{\delta}), \mathcal{M}(\mathbf{p}_{\mathcal{S}}(\mathbf{x}))\big). \tag{5}$$

Integrating adversaries into weak-to-strong generalization elicits robustness even with guidance from a non-robust weak VLM (Table 1). The necessity of employing adversaries arises from their ability to represent worst-case input distributions, enabling the strong student VLM to better mimic the weak teacher's robust behavior and its own self-knowledge refinement. Thus, we focus on adversary-driven weak-to-strong generalization for transferable robustness. To mitigate potentially unreliable supervision from weak teacher VLMs, we investigate two complementary mechanisms: *alignment re-weighting* and *source guidance refinement*, supported by theoretical analyses in the following sections.

3.3 Entropy-guided Uncertainty Re-weighting

Recall that in weak-to-strong generalization, the trade-off factor β balances teacher-student alignment and student self-refinement. However, β is either fixed or follows a predefined warm-up schedule, neglecting the potential prediction errors from the weak teacher model. Thus, we propose an adaptive re-weighting mechanism guided by the teacher's prediction uncertainty, quantified via entropy:

Teacher entropy. We define teacher prediction entropy as $\mathsf{H}_{\mathrm{T}}(\mathbf{x}) = -\sum_c \left[\mathbf{p}_{\mathcal{T}}(\mathbf{x})\right]_c \log\left[\mathbf{p}_{\mathcal{T}}(\mathbf{x})\right]_c$, capturing uncertainty in an unsupervised scheme: higher entropy $\mathsf{H}_{\mathrm{T}}(\mathbf{x})$ indicates greater uncertainty (predictions are close to uniform distributions), whereas low entropy reflects more confident predictions. By leveraging prediction entropy as a reliability quantification, we adaptively emphasize the low-entropy guidance from the teacher model, enhancing robust knowledge superalignment.

We seek an instance-wise re-weighting function $\mathsf{w}(\cdot)$ that maps prediction entropy to weights while preserving relative differences in uncertainty across samples. To avoid uncertainty scale distortions from ad hoc normalization, w is required to be monotonically increasing in H_{T} (so that more uncertain examples receive higher weight). A suitable choice is a soft normalization:

$$\mathsf{w}(\mathbf{x}) = \frac{\mathsf{H}_{\mathrm{T}}(\mathbf{x})}{\mathsf{H}_{\mathrm{T}}(\mathbf{x}) + \kappa_{\mathsf{H}}}, \tag{6}$$

where κ_{H} is obtained by an adaptive scaling strategy as the median entropy of the teacher's prediction over the dataset: $\kappa_{\mathsf{H}} = \mathrm{median}\left(\{\mathsf{H}_{\mathrm{T}}(\mathbf{x}) | \mathbf{x} \in \mathcal{D}\}\right)$. The median entropy serves as a robust central tendency measure against outliers, allowing confident teacher predictions to guide alignment, while uncertain cases prioritize the student self-refinement. Robust weak-to-strong generalization (Eq. (5)) can be reformulated by replacing fixed β with our entropy-guided weighting $\mathsf{w}(\mathbf{x})$ per input $\mathbf{x}$.

Theorem 1 (Adapted from [25])**.** *Let the weights of the classifier, i.e., in our case these are textual VLM embeddings denoted as $\boldsymbol{\psi}_c = f_{\boldsymbol{\theta}_T}(\mathbf{t}_c)$, and the vision feature extractor be given by $\phi(\cdot)$. Let $\mathsf{H}[\cdot]$ be the Shannon entropy. Then for the conditional entropy over predictions it holds true that:*

$$\|\phi(\mathbf{x})\|_2 \geq \frac{\log(C) - \mathsf{H}\left[\mathbf{p}(\cdot \mid \phi(\mathbf{x}), \boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_C)\right]}{2\max_{i=1,\ldots,C} \|\boldsymbol{\psi}_i\|_2}. \tag{7}$$

Proof. See [25]. □

Hence, with fixed textual embeddings (denominator), lower prediction entropy tightens the model selection space for the vision encoder in VLMs by favoring models with larger weights. The vision encoder is thus constrained to a more stable optimization trajectory and is less prone to overfitting.

Definition 1 (Classification margin in VLMs)**.** *Consider a VLM of the form $\left(\phi(\cdot), \{\boldsymbol{\psi}_c\}_{c=1}^C\right)$, where $\phi(\mathbf{x}) \in \mathbb{R}^d$ is the vision feature for input $\mathbf{x}$, and each $\boldsymbol{\psi}_c \in \mathbb{R}^d$ is a text embedding representing class c. For input $\mathbf{x}$ with ground-truth label y, the classification margin is defined as:*

$$\gamma(y|\mathbf{x}) := \boldsymbol{\psi}_y^\top \phi(\mathbf{x}) - \max_{c \neq y} \boldsymbol{\psi}_c^\top \phi(\mathbf{x}). \tag{8}$$

The input $\mathbf{x}$ is correctly classified as y when the classification margin $\gamma(y|\mathbf{x}) > 0$, with larger $\gamma(y|\mathbf{x})$ indicating a wider separation from the nearest decision boundary.

Theorem 2. *Let $\left(\phi(\cdot), \{\boldsymbol{\psi}_c\}_{c=1}^C\right)$ be the same VLM setup as in Definition 1. Assume $\phi(\cdot)$ is L-Lipschitz continuous w.r.t. the input norm, i.e., $\|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|_2 \leq L\|\mathbf{x} - \mathbf{x}'\|$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. Suppose input $\mathbf{x}$ satisfies Theorem 1 (i.e., low prediction entropy enforces large feature norm $\|\phi(\mathbf{x})\|_2$). Under the* mild alignment *assumption that $\phi(\mathbf{x})$ is not close-to-orthogonal to the ground-truth textual embedding $\boldsymbol{\psi}_y$ of class y, increasing the vision feature norm $\|\phi(\mathbf{x})\|_2$ expands the margin $\gamma(\mathbf{x})$. If $\gamma(\mathbf{x}) > 0$, the prediction for $(\mathbf{x} + \boldsymbol{\delta})$ cannot be altered away from the ground-truth label y for any perturbation $\boldsymbol{\delta} \in \mathcal{X}$ with:*

$$\|\delta\|_\infty \leq \|\delta\|_2 < \frac{\gamma(\mathbf{x})}{L\ \max_{c \neq y} \|\boldsymbol{\psi}_c - \boldsymbol{\psi}_y\|_2}. \tag{9}$$

Proof. See Appendix C.1. □

Combined Theorems 1 & 2, and Definition 1 indicate that *low-entropy prediction* enforces a large vision feature norm, which in turn *amplifies* the classification margin. Under a mild Lipschitz assumption, this margin translates to *robustness* against input perturbations. In other words, a model with highly confident predictions naturally places each sample far from the decision boundary in feature space, making it less susceptible to small adversarial or noisy modifications.

3.4 Teacher Guidance Refinement via Inverse Adversarial Examples

Beyond re-weighting uncertain teacher predictions, we investigate an adaptive prediction refinement scheme to reduce reliance on erroneous supervision. In other words, we focus on generating more benign inputs to improve guidance for robust weak-to-strong generalization model. Instead of generating adversaries that reduce confidence (*i.e.*, loss-input gradient ascent), we consider an inverse procedure by reversing the gradient to maximize the likelihood in the neighborhood region of clean samples. Thus, this input perturbation enhances teacher prediction confidence. Such refined samples highlight class-specific features, offering more reliable guidance for the robust weak-to-strong generalization framework.

Inverse adversarial perturbation formulation. Given an input $\mathbf{x}$, we aim to generate a benignly perturbed sample $\check{\mathbf{x}} = \mathbf{x} + \check{\boldsymbol{\delta}}$ within an ℓ_∞-bounded region ($\|\check{\boldsymbol{\delta}}\|_\infty \le \check{\epsilon}$), which maximizes the teacher's prediction confidence. Formally, this inverse adversarial example is defined by:

$$\check{\mathbf{x}}^{(i+1)} = \Pi_{\mathbb{B}(\mathbf{x},\check{\epsilon})}\left[\check{\mathbf{x}}^{(i)} - \alpha \cdot \operatorname{sign}\left(\nabla_{\check{\mathbf{x}}^{(i)}} \mathcal{L}_{\text{CE}}\big(\mathbf{p}_{\mathcal{T}}(\check{\mathbf{x}}^{(i)}), \mathcal{M}(\mathbf{p}_{\mathcal{T}}(\mathbf{x}))\big)\right)\right]. \tag{10}$$

Equivalently, the perturbation $\check{\boldsymbol{\delta}}$ shifts the teacher's prediction $\mathbf{p}_{\mathcal{T}}(\check{\mathbf{x}})$ closer to a one-hot distribution based on the teacher-derived pseudolabels $\mathcal{M}(\mathbf{p}_{\mathcal{T}}(\mathbf{x}))$. This process contrasts standard adversary generation Eq. (3) by reinforcing, rather than undermining, the teacher's initial decision, creating an "inverse adversary". Consequently, the resulting inverse adversary $\check{\mathbf{x}} = \mathbf{x} + \check{\boldsymbol{\delta}}$ enhances class-specific confidence and refines teacher supervision, supporting robust weak-to-strong generalization. We reformulate $\mathcal{L}_{\text{CE}}\big(\mathbf{p}_{\mathcal{S}}(\mathbf{x}+\boldsymbol{\delta}), \mathbf{p}_{\mathcal{T}}(\mathbf{x})\big)$ into its refined counterpart $\mathcal{L}_{\text{CE}}\big(\mathbf{p}_{\mathcal{S}}(\mathbf{x}+\boldsymbol{\delta}), \mathbf{p}_{\mathcal{T}}(\check{\mathbf{x}})\big)$ by aligning adversarial predictions with high-confidence teacher outputs. To justify this refinement, Theorem 3 formalizes the unique role of inverse adversaries as effective guidance distinct from standard inputs.

Theorem 3. *Let $\mathbf{p}_{\mathcal{T}}(\mathbf{x})$ and $\mathbf{p}_{\mathcal{S}}(\mathbf{x})$ denote the softmax predictions over C categories w.r.t. the* teacher *and* student *VLM during weak-to-strong generalization, respectively. Suppose $\hat{\mathbf{x}} = \mathbf{x} + \boldsymbol{\delta}$ is an* adversarial example *for student, while $\check{\mathbf{x}} = \mathbf{x} + \check{\boldsymbol{\delta}}$ is an* inverse *adversarial example for the teacher with high confidence in the ground-truth class. Then the following cross-entropy inequality holds:*

$$\mathcal{L}_{CE}\big(\mathbf{p}_{\mathcal{S}}(\hat{\mathbf{x}}), \mathbf{p}_{\mathcal{T}}(\check{\mathbf{x}})\big) \ge \mathcal{L}_{CE}\big(\mathbf{p}_{\mathcal{S}}(\mathbf{x}), \mathbf{p}_{\mathcal{T}}(\check{\mathbf{x}})\big), \tag{11}$$

where $\mathcal{L}_{CE}(\mathbf{q}, \mathbf{p}) = -\sum_{i=1}^{C} p_i \log(q_i)$ is the cross-entropy of $\mathbf{q}$ w.r.t. target $\mathbf{p}$.

Proof. See Appendix C.2. □

Theorem 3 shows a unique and beneficial property of using our proposed unsupervised inverse adversaries as guidance during robust weak-to-strong generalization. Reducing the prediction gap between $\mathbf{p}_{\mathcal{T}}(\check{\mathbf{x}})$ and $\mathbf{p}_{\mathcal{S}}(\hat{\mathbf{x}})$ also enhances the standard teacher-student prediction alignment for clean samples, thus enhancing natural generalization to unseen data.

Objective function. We formulate Adv-W2S as a simple min-max optimization by replacing the learning objectives in standard weak-to-strong generalization (Eq. (4)) with adversarial and inverse adversarial examples. Under the unsupervised setting, we conduct both adversarial perturbation $\boldsymbol{\delta}$ and inverse adversarial perturbation $\check{\boldsymbol{\delta}}$ generation based on the pseudolabels from the teacher VLM:

$$\boldsymbol{\delta} = \arg\max_{\|\boldsymbol{\delta}\|_\infty \le \epsilon} \mathcal{L}_{\text{CE}}\big(\mathbf{p}_{\mathcal{S}}(\mathbf{x}+\boldsymbol{\delta}), \mathbf{p}_{\mathcal{T}}(\mathbf{x})\big) \quad \text{and} \quad \check{\boldsymbol{\delta}} = \arg\min_{\|\check{\boldsymbol{\delta}}\|_\infty \le \check{\epsilon}} \mathcal{L}_{\text{CE}}\big(\mathbf{p}_{\mathcal{T}}(\mathbf{x}+\check{\boldsymbol{\delta}}), \mathcal{M}(\mathbf{p}_{\mathcal{T}}(\mathbf{x}))\big). \tag{12}$$

We can thus form adversarial example $\hat{\mathbf{x}} = \mathbf{x} + \boldsymbol{\delta}$ and its inverse adversarial counterpart $\check{\mathbf{x}} = \mathbf{x} + \check{\boldsymbol{\delta}}$ based on the maximization above. By reformulating the adversarial variant of weak-to-strong generalization (Eq. (5)) with our entropy-guided uncertainty re-weighting (Eq. (6)), our Adv-W2S minimizes:

$$\mathcal{L}^{\text{W-Inv}}_{\text{Adv-AuxConf}} = \big(1 - \mathsf{w}(\mathbf{x})\big) \cdot \mathcal{L}_{\text{CE}}\big(\mathbf{p}_{\mathcal{S}}(\hat{\mathbf{x}}), \mathbf{p}_{\mathcal{T}}(\check{\mathbf{x}})\big) + \mathsf{w}(\mathbf{x}) \cdot \mathcal{L}_{\text{CE}}\big(\mathbf{p}_{\mathcal{S}}(\hat{\mathbf{x}}), \mathcal{M}(\mathbf{p}_{\mathcal{S}}(\mathbf{x}))\big). \tag{13}$$

During the inference stage, we directly use the CLIP model fine-tuned via our Adv-W2S method for robustness evaluations and further cross-task generalization without modifying VLM architectures.

4 Experiments

In this section, we compare our Adv-W2S with state-of-the-art adversarial fine-tuning approach across various downstream tasks and scenarios. Below, we detail our experimental configurations.

Table 2: Zero-shot **clean** and **robust** accuracy (%). Adversarial learning is conducted on ImageNet with evaluations across 14 datasets. Adversaries are generated via Auto-Attack with radius $\epsilon = 2/255$.

Eval.	Method	ImageNet	STL10	CIFAR-10	CIFAR-100	Stanf.Cars	Caltech101	OxfordPet	Flower102	DTD	EuroSAT	FGVC	PCAM	ImageNet-R	ImageNet-S	Average
Clean Acc.	Standard CLIP	74.90	99.31	95.20	71.08	77.91	83.29	93.21	79.17	55.21	62.65	31.77	52.01	87.86	59.61	73.08
	TeCoA [55]	80.00	95.40	86.88	61.64	44.45	80.33	80.78	51.83	45.43	23.48	15.00	58.39	79.40	58.77	61.56
	PMG [69]	77.84	96.92	90.25	64.97	58.23	83.34	86.45	58.46	46.49	28.04	20.64	49.99	83.18	57.62	64.46
	FARE [63]	72.96	98.28	90.24	67.78	66.80	**85.65**	89.75	65.13	**50.43**	16.54	22.83	50.02	83.75	56.86	65.50
	TGA [71]	**80.26**	96.83	88.07	60.86	49.81	81.54	81.11	51.49	45.96	**30.30**	14.22	49.95	80.20	58.89	62.11
	Adv-W2S	75.84	**98.48**	**91.64**	**68.83**	**70.58**	84.86	**91.11**	**70.68**	49.57	24.96	**27.48**	**63.51**	**85.37**	**59.58**	**68.75**
Robust Acc.	Standard CLIP	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.01	0.01	0.09	0.01
	TeCoA [55]	**61.74**	86.34	61.99	35.82	18.62	70.57	68.22	27.27	26.17	**12.37**	5.43	26.93	59.57	44.56	43.26
	PMG [69]	60.02	88.21	64.12	37.14	23.68	72.47	70.92	28.20	26.33	9.07	5.79	47.06	62.24	45.08	45.74
	FARE [63]	43.56	88.55	61.82	34.89	23.74	70.88	67.70	32.95	25.69	3.76	5.31	**49.39**	56.47	36.86	42.97
	TGA [71]	61.46	88.56	63.21	35.44	21.60	71.16	68.52	26.15	26.70	11.37	5.76	47.88	60.32	44.46	45.19
	Adv-W2S	58.30	**90.49**	**69.56**	**41.08**	**29.62**	**73.03**	**73.24**	**34.33**	**29.63**	11.06	**7.53**	48.76	**65.32**	**45.43**	**48.38**

Table 3: Average accuracy (%) of diverse CLIP backbones with perturbation radius $\epsilon = 2/255$.

Backbone	Method	Clean	Robust
ResNet101	TeCoA [55]	43.40	23.40
	PMG [69]	46.34	22.48
	FARE [63]	47.88	16.61
	TGA [71]	45.29	23.18
	Adv-W2S	**50.14**	**25.93**
ViT-B/16	TeCoA [55]	49.46	33.41
	PMG [69]	53.57	33.05
	FARE [63]	54.09	30.54
	TGA [71]	51.32	33.07
	Adv-W2S	**55.42**	**35.20**

Table 4: Average accuracy (%) of diverse ϵ when fine-tuning and testing w.r.t. CLIP w/ ViT-L.

Radius	Method	Clean	Robust
$\epsilon = 3/255$	TeCoA [55]	58.90	38.07
	PMG [69]	61.72	39.40
	FARE [63]	63.55	37.17
	TGA [71]	59.65	38.59
	Adv-W2S	**64.79**	**40.85**
$\epsilon = 4/255$	TeCoA [55]	56.25	32.53
	PMG [69]	58.82	33.87
	FARE [63]	60.26	32.02
	TGA [71]	56.76	32.86
	Adv-W2S	**61.56**	**35.19**

Datasets. In line with prior works [55, 63], we conduct adversarial learning on the ImageNet training set [13], with zero-shot classification evaluations on its test set and other 13 datasets. We further explore downstream task generalization across various datasets w.r.t. image captioning, visual question answering, object hallucination, and science question answering (see Appendix B.1).

Implementation details. Unless specified otherwise, we adopt CLIP [60] with the ViT-Large/14 architecture, as in previous studies [55, 63]. During adversarial weak-to-strong generalization, we conduct adversary generation via 10-step PGD with perturbation radius $\epsilon = 2/255$ and step size $\alpha = 1/255$ in an unsupervised scheme. In analogy to superalignment, we consider a relatively weak teacher of the ViT-Base/32 architecture, pre-trained on the ImageNet training set using TeCoA [55]. Zero-shot robustness is assessed under Auto-Attack [12], an ensemble adversarial attack method for reliable evaluations. We achieve downstream task generalization based on two large-scale VLM frameworks, LLaVA 1.5 7B [48] and OpenFlamingo 9B [2], by replacing vision encoders with our robust counterparts. For fairness, evaluations are under adaptive attacks. Details are in Appendix B.2.

4.1 Main Results (Zero-shot Classification)

Zero-shot classification performance. Table 2 compares Adv-W2S with state-of-the-art adversarial fine-tuning methods using CLIP ViT-L. We report both clean and robust accuracy (Auto-Attack [12]) across 14 datasets. Our Adv-W2S achieves the best zero-shot performance with an average improvement of ~5% in clean accuracy and ~5.8% in robustness. While in-distribution accuracy on ImageNet drops slightly due to the unsupervised learning scheme [63], Section 4.3 shows this gap can be mitigated by incorporating a supervised objective into robust weak-to-strong generalization.

Robustness across diverse CLIP backbones. In addition to robustness with ViT-L, we here apply our Adv-W2S method on ResNet101 and ViT-Base/16 based on the weak teacher VLM of ResNet50 and ViT-Base/32, respectively. Table 3 shows that our Adv-W2S consistently outperforms other adversarial learning approaches in both average clean and robust accuracy across 14 datasets.

Adversarial learning with diverse perturbation radii. Beyond the default perturbation radius $\epsilon = 2/255$, we explore adversaries of larger ℓ_∞-norm perturbation radii ($\epsilon = 3/255, 4/255$) during both fine-tuning and robustness evaluations for fair comparisons. As shown in Table 4, we observe that our Adv-W2S surpasses other methods across diverse perturbation radii in zero-shot scenarios.

Table 6: Zero-shot transfer performance on image captioning (CIDEr score) and VQA (accuracy %).

VLM Type	Method	Captioning				Visual Question Answering					
		COCO		Flickr30k		TextVQA		VQAv2		Vizwiz	
		Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust
LLaVA 1.5	Standard CLIP	112.3	2.9	74.7	1.0	34.8	0.0	74.5	0.0	39.4	2.3
	TeCoA [55]	96.7	45.1	55.2	24.0	23.8	12.8	66.2	35.7	42.5	29.6
	PMG [69]	103.1	52.8	63.2	28.4	27.6	14.0	68.4	35.1	41.0	27.6
	FARE [63]	108.5	47.9	67.4	24.5	30.5	14.7	70.3	34.5	41.9	25.3
	TGA [71]	101.3	50.6	61.9	27.8	27.1	14.6	67.3	35.0	42.8	28.0
	Adv-W2S	**110.8**	**55.3**	**72.6**	**30.7**	**32.5**	**16.4**	**72.9**	**37.3**	**45.3**	**33.8**
OpenFlamingo	Standard CLIP	78.8	1.5	58.7	0.6	22.3	0.0	47.7	0.0	17.7	3.3
	TeCoA [55]	73.0	29.6	47.4	13.7	17.3	2.4	46.1	23.8	17.6	4.0
	PMG [69]	76.2	31.2	52.0	16.5	17.7	2.8	47.0	24.0	16.9	4.2
	FARE [63]	77.9	32.7	53.5	15.9	18.8	2.2	46.7	21.8	17.2	3.8
	TGA [71]	74.2	30.5	51.8	16.0	19.0	2.7	46.2	23.6	18.0	2.5
	Adv-W2S	**80.0**	**34.3**	**56.9**	**17.1**	**20.5**	**3.9**	**47.4**	**25.3**	**18.2**	**5.3**

Extension with adversarial PEFT. Fully fine-tuning large-scale VLMs typically introduces significant computational costs. We therefore extend adversarial VLM learning with LoRA [35], a PEFT strategy using trainable low-rank matrices for efficient adaptation. We report both clean and robust accuracy in the zero-shot setting of our LoRA-based Adv-W2S as well as other approaches with the LoRA strategy (see Table 5). The results indicate that our Adv-W2S can still outperform other approaches, even with the LoRA strategy for efficiency.

Table 5: Acc. (%) w.r.t. diff. ϵ for fine-tuning and testing (ViT-L) with LoRA.

Radius	Method	Clean	Robust
$\epsilon = 3/255$	TeCoA [55]	55.22	26.54
	PMG [69]	56.82	27.02
	FARE [63]	57.84	23.85
	TGA [71]	55.88	26.66
	Adv-W2S	**59.21**	**29.13**
$\epsilon = 4/255$	TeCoA [55]	49.84	19.87
	PMG [69]	52.08	20.07
	FARE [63]	53.20	19.09
	TGA [71]	51.13	19.83
	Adv-W2S	**54.28**	**21.91**

4.2 Zero-Shot Downstream Task Generalization

Image captioning extension. We here extend Adv-W2S to image captioning by replacing the vision encoders of large-scale VLMs (*e.g.*, LLaVA and OpenFlamingo) with our robust versions. We report the CIDEr score [66] for both COCO and Flickr30k datasets (see Table 6). We observe that our Adv-W2S achieves the best captioning performance in terms of both clean and adversarial examples compared to other adversarial learning approaches. In addition to quantitative results, we further provide image captioning visualizations against unforeseen adversaries in Figure 1 in Appendix F.

Visual Question Answering (VQA) extension. Table 6 reports VQA accuracy [1] across three standard VQA datasets using different VLMs. Our Adv-W2S method receives a large gain in zero-shot robustness while maintaining comparable natural performance with standard CLIP. Note that our method can even outperform standard CLIP in clean accuracy on Vizwiz, leading to lossless robustness enhancement. The corresponding visualizations are provided in Figure 2 in Appendix F.

Object hallucination extension. Foundation VLMs are vulnerable to object hallucinations (*i.e.*, erroneously recognizing objects that do not exist in inputs) [61]. *POPE* [44] served as an object hallucination evaluation benchmark based on VQA across diverse question sampling scenarios (details in Appendix B.3). We extend adversarial VLM learning with hallucination evaluations in Table 7 (see also visualizations in Figure 3 in Appendix F). Notably, our Adv-W2S shows reduced hallucination rates, benefiting from our inherent regularization to over-confident or overly aligned predictions by balancing teacher-student alignment with model uncertainty.

Table 7: Hallucination evaluations (F1-score) using POPE for adversarial VLM learning (ViT-L).

Method	POPE Sampling			Avg. Score
	Random	Popular	Adversarial	
TeCoA [55]	79.8	79.1	75.2	78.0
PMG [69]	81.7	80.9	76.3	79.6
FARE [63]	82.2	81.5	78.6	80.8
TGA [71]	80.4	79.8	76.0	78.7
Adv-W2S	85.6	84.9	81.0	83.8

Science question answering extension w/ CoT. Chain of Thought (CoT) has been widely explored to elicit intermediate reasoning steps from large-scale VLMs by guiding them to generate multi-step rationales before producing a final answer [8]. Science Question Answering [52] has been recognized as a standard evaluation benchmark for CoT due to its large quantity of multi-option questions with multimodal contexts from the science curriculum. As shown in Table 8, we extend diverse adversarial VLM

Table 8: CoT eval. (Acc.) using science question answering for adversarial VLM learning (ViT-L).

Setting	Method	Temperature			Avg. Acc.
		0.0	0.1	0.2	
Standard	TeCoA [55]	51.4	51.6	50.0	51.0
	PMG [69]	51.9	52.0	51.6	51.8
	FARE [63]	52.5	52.2	52.4	52.4
	TGA [71]	52.1	51.9	51.8	51.9
	Adv-W2S	**53.8**	**53.9**	**53.6**	**53.8**
Single Choice	TeCoA [55]	67.4	67.6	67.1	67.4
	PMG [69]	68.2	68.0	67.7	68.0
	FARE [63]	68.1	67.9	67.7	67.9
	TGA [71]	67.5	67.8	67.4	67.6
	Adv-W2S	**69.3**	**69.5**	**69.0**	**69.3**

learning methods to science question answering across different settings to evaluate their CoT capability. Further details of diverse configurations are in Appendix B.3. Our method consistently achieves the best science question answering performance across diverse settings, which demonstrates that our robust weak-to-strong generalization potentially enhances the CoT capability of large-scale VLMs. Visual examples are in Figure 4 in Appendix F.

4.3 Further Analyses (Why Adv-W2S is Effective)

Below we conduct systematic analyses of our proposed Adv-W2S framework and its component modules to justify its efficacy and generalization capability across diverse configurations.

Impact of component modules. We analyze the contribution of two main components of our Adv-W2S: (i) Entropy-guided Uncertainty Re-weighting (EUR) from Eq. (6), and (ii) Inverse Adversarial Refinement (IAR) in Eq. (10). Table 9 reports the average clean and robust accuracy across 14 datasets in the zero-shot setting. We adopt the adversarial variant of weak-to-strong generalization from Eq. (5) as the baseline (first row in Table 9). Enforcing instance-wise re-weighting to balance prediction alignment and self-refinement contributes to improving both natural performance and adversarial robustness. Refining the teacher guidance with inverse adversarial examples also enhances natural generalization to unseen data, yielding further gains in the zero-shot accuracy.

Table 9: Ablation study of key components in our Adv-W2S for average clean and robust accuracy (%) on 14 datasets.

	EUR	IAR	Clean	Robust
1			63.45	42.43
2	✓		66.42	47.35
3		✓	67.09	46.79
4	✓	✓	68.75	48.38

Impact of teacher-student setups. In the standard weak-to-strong generalization paradigm, the weak teacher model is typically pre-trained using task-specific supervised learning under natural (non-adversarial) conditions. To provide further insights into the teacher-student configurations employed in our adversarial weak-to-strong generalization (Adv-W2S) framework, we summarize the key setups (i) whether the weak teacher is pre-trained under adversarial or natural conditions, and (ii) whether an unsupervised adversarial fine-tuning (i.e., FARE [63]) is employed during the initial warm-up stage of Adv-W2S. Table 10 shows that an adversarially pre-trained teacher with unsupervised adversarial fine-tuning for student warm-up enjoys greater zero-shot robustness.

Table 10: Average performance (%) of our Adv-W2S method with diverse teacher-student setups.

Teacher	Student	Clean	Robust
Nat. Pre-Trained	w/o Adv. Warm-up	68.05	46.79
Adv. Pre-Trained	w/o Adv. Warm-up	67.53	47.40
Nat. Pre-Trained	w/ Adv. Warm-up	68.98	47.02
Adv. Pre-Trained	w/ Adv. Warm-up	68.75	48.38

Impact of adversary generation schemes. Below, we analyze the average zero-shot classification performance of diverse adversary generation schemes (*i.e.*, objective functions for generating adversarial examples $\hat{\mathbf{x}} = \mathbf{x} + \boldsymbol{\delta}$) in Table 11. We observe that adversary generation by maximizing the teacher-student prediction gap enforces a better robustness transfer in the context of weak-to-strong generalization. More details are in Appendix D.1.

Table 11: Average accuracy (%) of our Adv-W2S with diverse **adversary generation** schemes.

Adversary Generation	Clean	Robust
Student Feature Deviation	67.17	46.69
Student Prediction Deviation	65.98	47.20
Teacher-Student Prediction Distance	68.75	48.38

Impact of inverse adversary generation schemes. In addition to diverse adversary generation strategies, we also investigate a range of inverse adversary generation approaches (details in Appendix D.2) in Table 12. The results indicate that using the pseudolabel cross-entropy loss leads to inverse adversaries of more benign guidance for robust weak-to-strong generalization.

Table 12: Average accuracy (%) of Adv-W2S with diverse **inverse adversary generation** schemes.

Inverse Adversary Generation	Clean	Robust
Pseudo-Margin Minimization	68.19	47.82
Prediction Entropy Minimization	67.93	47.34
Pseudolabel Cross-Entropy	68.75	48.38

Auxiliary ground-truth supervision. Recall that we primarily focus on robust weak-to-strong generalization in an unsupervised scheme following the standard setup [5]. Despite its improved zero-shot performance, its in-distribution performance on the fine-tuned dataset is still lower than supervised adversarial fine-tuning at the same VLM backbone. Thus, we investigate the underlying effect

Table 13: Clean and robust accuracy (%) of our Adv-W2S with/without an **auxiliary ground-truth supervision**.

Method	ImageNet		Avg. 13 Datasets	
	Clean	Robust	Clean	Robust
TeCoA [55]	80.00	61.74	60.14	41.84
Unsupervised **Adv-W2S**	75.84	58.30	**68.20**	**47.62**
Supervised **Adv-W2S**	**80.56**	**64.08**	65.72	45.25

of appending an auxiliary ground-truth supervision in our Adv-W2S method (see Table 13). Details of this auxiliary branch are in Appendix E. We observe an inherent trade-off between in-distribution performance on ImageNet and out-of-distribution (zero-shot) performance on other datasets.

Hyper-parameter sensitivity analyses We further provide a systematic analysis of key hyperparameters involved in our proposed Adv-W2S method in Appendix G.

5 Conclusions

Motivated by our analysis of standard weak-to-strong generalization paradigm in eliciting VLM robustness, we have uncovered that neglecting adversarial examples in alignment objectives leads to robustness degradation, even when leveraging source guidance from an adversarially pre-trained VLM. Thus, we have investigated an adversarial adaptation of standard weak-to-strong generalization, explicitly integrating adversarial examples to elicit robust knowledge in an unsupervised scheme. Recognizing that supervision from a weak-capacity VLM may be inherently unreliable, we introduce two complementary strategies: uncertainty-based alignment re-weighting and source guidance refinement via inverse adversarial examples. Further theoretical analyses characterize the robustness elicitation efficacy of our method, demonstrating enhanced generalization against subtle adversarial or noise modifications.

Acknowledgments

This research is supported in part by National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative, partly supported through the AI Singapore Programme under the project titled “AI-based Urban Cooling Technology Development”(Award No. AISG3-TC-2024-014-SGKR), the Centre for Frontier Artificial Intelligence Research, Institute of High Performance Computing, A*Star, and the College of Computing and Data Science at Nanyang Technological University. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore, and Infocomm Media Development Authority. Piotr Koniusz is supported by CSIRO’s Science Digital.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 8, 18

[2] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 3, 7, 18

[3] Rishika Bhagwatkar, Shravan Nayak, Pouya Bashivan, and Irina Rish. Improving adversarial robustness in vision-language models with architecture and prompt design. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17003–17020, 2024. 3

[4] Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, et al. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*, 2022. 1

[5] Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeffrey Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *Forty-first International Conference on Machine Learning, ICML*, 2024. 1, 3, 4, 9

[6] Paul H Calamai and Jorge J Moré. Projected gradient methods for linearly constrained problems. *Mathematical programming*, 39(1):93–116, 1987. 3

[7] Paul Christiano, Buck Shlegeris, and Dario Amodei. Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*, 2018. 1

[8] Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. Navigate through enigmatic labyrinth A survey of chain of thought reasoning: Advances, frontiers and future. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*, pages 1173–1203, 2024. 8

[9] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 17

[10] 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*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011. 17

[11] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. 4

[12] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020. 4, 7, 18

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 7, 17

[14] Junhao Dong, Piotr Koniusz, Junxi Chen, and Yew-Soon Ong. Adversarially robust distillation by reducing the student-teacher variance gap. In *European Conference on Computer Vision*, pages 92–111. Springer, 2024. 2

[15] Junhao Dong, Piotr Koniusz, Junxi Chen, Z Jane Wang, and Yew-Soon Ong. Robust distillation via untargeted and targeted intermediate adversarial samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28432–28442, 2024. 2

[16] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21037–21047, October 2025. 3

[17] Junhao Dong, Piotr Koniusz, Xinghua Qu, and Yew-Soon Ong. Stabilizing modality gap & lowering gradient norms improve zero-shot adversarial robustness of vlms. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, KDD '25, page 236–247, New York, NY, USA, 2025. Association for Computing Machinery. 3

[18] 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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 14061–14078. PMLR, 13–19 Jul 2025. 3

[19] Junhao Dong, Jiao Liu, Xinghua Qu, and Yew-Soon Ong. Confound from all sides, distill with resilience: Multi-objective adversarial paths to zero-shot robustness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 624–634, October 2025. 3

[20] Junhao Dong, Seyed-Mohsen Moosavi-Dezfooli, Jianhuang Lai, and Xiaohua Xie. The enemy of my enemy is my friend: Exploring inverse adversaries for improving adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24678–24687, June 2023. 2

[21] Junhao Dong, Yuan Wang, Jian-Huang Lai, and Xiaohua Xie. Improving adversarially robust few-shot image classification with generalizable representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9025–9034, 2022. 2

[22] Junhao Dong, Yuan Wang, Xiaohua Xie, Jianhuang Lai, and Yew-Soon Ong. Generalizable and discriminative representations for adversarially robust few-shot learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3):5480–5493, 2024. 2

[23] Junhao Dong, Lingxiao Yang, Yuan Wang, Xiaohua Xie, and Jianhuang Lai. Toward intrinsic adversarial robustness through probabilistic training. *IEEE Transactions on Image Processing*, 32:3862–3872, 2023. 2

[24] Junhao Dong, Hao Zhu, Yifei Zhang, Xinghua Qu, Yew-Soon Ong, and Piotr Koniusz. Machine unlearning via task simplex arithmetic. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1

[25] Abhimanyu Dubey, Otkrist Gupta, Ramesh Raskar, and Nikhil Naik. Maximum-entropy fine grained classification. *Advances in neural information processing systems*, 31, 2018. 5

[26] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 17

[27] Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3996–4003, 2020. 4

[28] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 17

[29] Jianyuan Guo, Hanting Chen, Chengcheng Wang, Kai Han, Chang Xu, and Yunhe Wang. Vision superalignment: Weak-to-strong generalization for vision foundation models. *arXiv preprint arXiv:2402.03749*, 2024. 3

[30] Yue Guo and Yi Yang. Improving weak-to-strong generalization with reliability-aware alignment. *arXiv preprint arXiv:2406.19032*, 2024. 3

[31] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 17

[32] 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. 17

[33] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021. 17

[34] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

[35] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR*, 2022. 3, 8, 18

[36] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 3

[37] Jaehyung Kim, Yuning Mao, Rui Hou, Hanchao Yu, Davis Liang, Pascale Fung, Qifan Wang, Fuli Feng, Lifu Huang, and Madian Khabsa. Roast: Robustifying language models via adversarial perturbation with selective training. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 3412–3444, 2023. 3

[38] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 17

[39] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 17

[40] Hunter Lang, David Sontag, and Aravindan Vijayaraghavan. Theoretical analysis of weak-to-strong generalization. *Advances in neural information processing systems*, 37:46837–46880, 2024. 3

[41] Jan Leike and Ilya Sutskever. Introducing Superalignment. *OpenAI Blog*, 2023. 1, 25

[42] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1

[43] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 3

[44] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 292–305, 2023. 8, 18, 24

[45] 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 *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 17

[46] Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Yu Cheng, and Wei Hu. A survey of attacks on large vision-language models: Resources, advances, and future trends. *arXiv preprint arXiv:2407.07403*, 2024. 2

[47] Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Xiang Fang, Keke Tang, Yao Wan, and Lichao Sun. Pandora's box: Towards building universal attackers against real-world large vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2

[48] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 3, 7, 18

[49] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1, 3

[50] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations, ICLR*, 2019. 17

[51] Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 102–111, 2023. 2

[52] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 8, 17, 24

[53] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR*, 2018. 17

[54] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 17

[55] Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. In *The Eleventh International Conference on Learning Representations,ICLR*, 2023. 3, 4, 7, 8, 9, 16, 17, 23

[56] Yao Ni, Shan Zhang, and Piotr Koniusz. PACE: marrying the generalization of PArameter-efficient fine-tuning with consistency regularization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3

[57] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 17

[58] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 17

[59] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 17

[60] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 7, 16, 17

[61] Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. A comprehensive survey of hallucination in large language, image, video and audio foundation models. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 11709–11724, 2024. 8

[62] Christian Schlarmann and Matthias Hein. On the adversarial robustness of multi-modal foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop*, pages 3677–3685, 2023. 18

[63] Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In *International Conference on Machine Learning*, pages 43685–43704. PMLR, 2024. 3, 7, 8, 9, 16, 17, 18

[64] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 17

[65] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014. 2

[66] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 8, 18

[67] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pages 210–218. Springer, 2018. 17

[68] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019. 17

[69] Sibo Wang, Jie Zhang, Zheng Yuan, and Shiguang Shan. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 7, 8, 16

[70] Xilie Xu, Keyi Kong, Ning Liu, Lizhen Cui, Di Wang, Jingfeng Zhang, and Mohan S. Kankanhalli. An LLM can fool itself: A prompt-based adversarial attack. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, 2024. 3

[71] Lu Yu, Haiyang Zhang, and Changsheng Xu. Text-guided attention is all you need for zero-shot robustness in vision-language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2024. 7, 8, 17

[72] Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pretraining models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5005–5013, 2022. 2

[73] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[74] Yifei Zhang, Hao Zhu, Junhao Dong, Haoran Shi, Ziqiao Meng, and Han Yu Piotr Koniusz. CrossSpectra: exploiting cross-layer smoothness for parameter-efficient fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 3

[75] Yudong Zhang, Ruobing Xie, Jiansheng Chen, Xingwu Sun, and Yu Wang. Pip: Detecting adversarial examples in large vision-language models via attention patterns of irrelevant probe questions. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11175–11183, 2024. 3

[76] Hao Zhu, Yifei Zhang, Junhao Dong, and Piotr Koniusz. BiLoRA: almost-orthogonal parameter spaces for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25613–25622, June 2025. 3

Robust SuperAlignment: Weak-to-Strong Robustness Generalization for Vision-Language Models (*Supplementary Material*)

Junhao Dong[1,2], Cong Zhang[3], Xinghua Qu[4], Zejun Ma[3], Piotr Koniusz[5,6,7], and Yew-Soon Ong[1,2]
[1]Nanyang Technological University, [2]CFAR, IHPC, A*STAR, [3]TikTok, Singapore, [4]Bytedance, [5]Data61♥CSIRO, [6]University of New South Wales, [7]Australian National University

Abstract

This supplementary material provides additional theoretical, algorithmic, and empirical details to support our main paper. We first present a comprehensive overview of adversarial fine-tuning schemes in vision-language models (Appendix A) and detail our experimental configurations, including dataset descriptions, implementation details, and downstream task extensions such as image captioning, visual question answering, object hallucination, and scientific reasoning (Appendix B). Appendix C offers formal proofs for our theoretical statements. We further elaborate on diverse objective functions for adversary and inverse adversary generation (Appendix D). Details with respect to the auxiliary supervised branch are provided in Appendix E, and the hyperparameter sensitivity analyses are in Appendix G. Finally, visualizations and extended qualitative results are provided in Appendix F, and broader impact and limitations are discussed in Appendix H.

A Further Adversarial Fine-tuning Schemes

We have analyzed the standard adversarial fine-tuning scheme (TeCoA) [55] in Eq. (2), and we here provide more details regarding other adversarial fine-tuning schemes in the context of Vision-Language Models (VLMs) for a more comprehensive background introduction.

PMG [69]. Motivated by the inherent overfitting of TeCoA [55] with generalization degradation, *PMG* [69] leveraged the prediction-level guidance from the vanilla pre-trained VLM with a regularization of clean samples for the target model to conduct adversarial fine-tuning, as follows:

$$\min_{\boldsymbol{\theta}_{\mathrm{I}}} \mathbb{E}_{(\mathbf{x},c)\sim\mathcal{D}}\Big[\max_{\|\boldsymbol{\delta}\|_\infty\leq\epsilon} \mathcal{L}_{\mathrm{CE}}\big(\mathbf{p}(\mathbf{x}+\boldsymbol{\delta}),\mathbf{y}(c)\big)+\lambda_1\cdot\mathcal{L}_{\mathrm{KL}}\big(\mathbf{p}_{\mathrm{orig}}(\mathbf{x})\|\mathbf{p}(\mathbf{x}+\boldsymbol{\delta})\big)+\lambda_2\cdot\mathcal{L}_{\mathrm{KL}}\big(\mathbf{p}(\mathbf{x})\|\mathbf{p}(\mathbf{x}+\boldsymbol{\delta})\big)\Big], \quad (14)$$

where $\mathcal{L}_{\mathrm{KL}}$ represents the Kullback–Leibler divergence, and $\mathbf{p}_{\mathrm{orig}}$ denotes the prediction of the vanilla pre-trained CLIP model [60]. λ_1 and λ_2 are the corresponding loss weighting factors.

FARE [63]. To enhance the robustness generalization capability across diverse vision-language tasks, Schlarmann *et al.* [63] proposed an unsupervised adversarial fine-tuning approach, dubbed *FARE*, to adversarially optimize feature-level discrepancies in an unsupervised scheme:

$$\min_{\boldsymbol{\theta}_{\mathrm{I}}} \mathbb{E}_{(\mathbf{x},c)\sim\mathcal{D}}\left[\max_{\|\boldsymbol{\delta}\|_\infty\leq\epsilon}\left\|f_{\mathrm{orig}}(\mathbf{x})-f(\mathbf{x}+\boldsymbol{\delta})\right\|_2^2\right], \quad (15)$$

where $f(\cdot)$ denotes the image encoder of the CLIP model for fine-tuning, while $f_{\mathrm{orig}}(\cdot)$ is the image encoder of the vanilla pre-trained CLIP model as the frozen reference.

*The research work was done during Junhao's internship at TikTok Singapore.
†Corresponding author.

TGA [71]. Motivated by the empirical observation that adversarial perturbations induce shifts in text attention maps, TGA [71] has been proposed to incorporate test-guided attention into adversarial fine-tuning, improving the zero-shot adversarial robustness of VLMs:

$$\min_{\boldsymbol{\theta}_{\mathrm{I}}} \mathbb{E}_{(\mathbf{x},c)\sim\mathcal{D}}\bigg[\max_{\|\boldsymbol{\delta}\|_{\infty}\leq\epsilon} \mathcal{L}_{\mathrm{CE}}\big(\mathbf{p}(\mathbf{x}+\boldsymbol{\delta}), \mathbf{y}(c)\big) + \lambda_1 \cdot \big\|g_{\mathrm{orig}}(\mathbf{x}) - g(\mathbf{x}+\boldsymbol{\delta})\big\| + \lambda_2 \cdot \big\|g_{\mathrm{orig}}(\mathbf{x}) - g(\mathbf{x})\big\|\bigg], \quad (16)$$

where $g(\cdot)$ extracts the text-guided attention map. While the original TGA mechanism relies on the global `[CLS]` token available in ViT-based CLIP models to compute text-guided attention maps, this approach is not directly applicable to ResNet-based CLIP architectures due to the absence of an explicit global pooling token. To address this limitation, we adapt the TGA computation by operating on the final spatial feature maps extracted from the ResNet-101 backbone (as shown in Table 3).

B Experimental Configurations

In this section, we present a comprehensive overview of the experimental setup for our Adv-W2S framework, including dataset descriptions, implementation details, and the methodology for extending our approach to diverse vision-language downstream tasks.

B.1 Dataset Description

In line with prior adversarial VLM learning works [55, 63], we conduct adversarial weak-to-strong generalization on the training set of ImageNet [13] and evaluate its classification performance on the validation set of ImageNet, as the ground-truth labels for the test set are not publicly available. We further evaluate the zero-shot classification performance across a broad spectrum of image recognition tasks, encompassing 13 datasets:

- **Natural Object Recognition:** STL-10 [10], CIFAR-10/100 [39], and Caltech-101 [26].
- **Fine-Grained Recognition:** Stanf. Cars (Stanford Cars) [38], Oxford-IIIT Pets [58], Flower102 [57], and FGVC Aircraft [54].
- **Texture Recognition:** DTD (Describable Textures Dataset) [9].
- **Remote Sensing Classification:** EuroSAT [32].
- **Medical Image Diagnosis:** PCAM (PatchCAMelyon) [67].
- **Robust Classification:** ImageNet-R(endition) [33] and ImageNet-S(ketch) [68].

In addition to zero-shot image classification, we further evaluate zero-shot transfer performance on a range of image-text downstream tasks, including image captioning and Visual Question Answering (VQA), assessing both natural performance and adversarial robustness. For image captioning, evaluations are conducted on the COCO [45] and Flickr30k [59] datasets. For VQA, we assess performance on TextVQA [64], VQAv2 [28], and VizWiz [31]. Object hallucination is evaluated on the COCO dataset [45], while Chain-of-Thought (CoT) reasoning is assessed on the SCIENCE QA benchmark [52], a large-scale multi-choice dataset comprising multimodal science questions with detailed explanations and diverse domain coverage.

B.2 Implementation Details (Zero-shot Classification)

Standard setups. In line with the configurations used in prior studies [55, 63], we adopt the CLIP model [60] with the ViT-Large/14 architecture for the strong student model, unless stated otherwise. In analogy to superalignment, we consider adversarial weak-to-strong generalization from a weak-capacity teacher CLIP model of the ViT-Base/32 architecture that is adversarially pre-trained on the ImageNet training set using TeCoA [55]. For robust weak-to-strong generalization of other CLIP architectures (see Table 3 in the main text), we consider the relatively strong-capacity student model of ResNet101 and ViT-Base/16 with the corresponding weak-capacity teacher model of ResNet50 and ViT-Base/32, respectively. During adversarial weak-to-strong generalization, we conduct adversary generation via 10-step PGD [53] with perturbation radius $\epsilon = 2/255$ and step size $\alpha = 1/255$ in an unsupervised scheme (see Eq. (12)). We set the inverse adversarial perturbation radius as $\check{\epsilon} = 2/255$. For network parameter optimization, we adopt the AdamW optimizer [50] with momentum coefficients $(0.9, 0.95)$. The adversarial weak-to-strong generalization is optimized with a cosine annealing learning rate schedule with a linear warm-up to the maximum learning rate of 1×10^{-5} for 2 epochs. During the warm-up period, we also integrate the objective function of FARE [63] (See Appendix A) with a tiny weighting factor of $\lambda_{\text{warm-up}} = 0.2$. For Parameter-Efficient Fine-Tuning (PEFT) extension of our Adv-W2S method in Table 5, we adopt the Low-Rank Adaptation (LoRA) [35] framework, applied specifically to the attention modules.

Evaluation protocol for zero-shot classification. Following [63], we evaluate the classification performance in terms of clean samples and their adversarial counterparts. Specifically for adversary generation during evaluations, we adopt Auto-Attack (AA) [12] with the maximum perturbation radius of $\epsilon = 2/255$ unless specified otherwise. For fairness, all the robustness evaluations are conducted under adaptive attacks. Note that zero-shot classification with CLIP is performed by computing the cosine similarity between image embeddings and a set of text embeddings corresponding to category-specific prompts. The input image is then classified by choosing the label whose text embedding has the highest similarity to the input image representation.

B.3 Downstream Task Extensions

We achieve downstream task generalization based on two large-scale VLM frameworks, LLaVA 1.5 7B [48] and OpenFlamingo 9B [2], by replacing their vision encoders of the ViT-Large/14 architecture with our robust counterparts. We further explain specific modifications and evaluation protocols for individual vision-language task extensions below.

Configurations for image captioning extensions. The CIDEr score [66] is adopted to measure the similarity of a generated sentence against a set of ground-truth sentences for image captioning evaluations in this paper. For adversary generation in the context of image captioning, we conduct APGD attacks [12] with 100 steps against each ground-truth caption w.r.t. each image following [63]. After each attack step, we compute the CIDEr scores and exclude samples from further attacks if their scores fall below our predefined thresholds, 10 for COCO and 2 for Flickr30k, representing less than 10% of the baseline (LLaVA) performance. This selective strategy allows us to allocate computational resources more effectively by concentrating on samples that retain relatively high scores. In the final stage, we conduct an adversarial attack with the perturbation that previously resulted in the lowest CIDEr score and using the corresponding ground-truth reference.

Configurations for visual question answering extensions. For VQA tasks, we report VQA accuracy [1] and adopt a similar evaluation scheme to image captioning, without using a prediction score threshold. Note that we select the five most frequent ground-truth answers from the ten available annotations. We further conduct targeted adversarial attacks based on the text prompt strings "Maybe" and "Word". Such an evaluation protocol is also aligned with [62, 63] with APGD attacks.

Configurations for object hallucination extensions. For object hallucination evaluations, we focus on the POPE benchmark [44] to convert the standard hallucination evaluation into a binary classification task by prompting LVLMs with basic yes-or-no short questions about the probing objects in three object sampling scenarios:

- **Random sampling:** Objects absent from the image are selected at random.
- **Popular sampling:** From the set of objects not present in the current image, we choose the top-k most common objects across the entire dataset.
- **Adversarial sampling:** We rank all candidate objects by their co-occurrence frequency with ground-truth objects and choose the top-k among those not appearing in the input image.

Configurations for chain of thought extensions. We assess on the ScienceQA test split based on the LLaVA framework. During inference, we follow two prompting/evaluation regimes: **standard** and **single-choice**. In the standard setting, the target VLM first produces an unrestricted chain-of-thought (CoT) rationale and is then re-prompted to extract its final answer token, emulating free-form scientific reasoning. While in the single-choice setting, we append the instruction `''Answer with the option's letter from the given choices directly.''` to the initial prompt, forcing the model to emit exactly one of {A, B, C, D}. We then test both scenarios w.r.t. different temperatures. Note that in Table 8 in the main text, the single-choice setup typically achieves better performance than the standard case. Constraining the decoder to output only the option label can potentially eliminate a major error source present in the standard CoT pipeline. The classifier-style prompt concentrates probability mass on four symbols, reducing exposure to sampling noise and temperature-induced drift that can corrupt the final answer token.

C Theoretical Analyses

C.1 Proof of Theorem 2

We now detail the proof of Theorem 2 that a large feature-space norm can yield a large classification margin, which in turn leads to *robustness against small input perturbation*.

Theorem 4 (Theorem 2 from the main text)**.** *Let $\left(\phi(\cdot), \{\psi_c\}_{c=1}^{C}\right)$ be the same VLM setup as in Definition 1. Assume $\phi(\cdot)$ is L-Lipschitz continuous w.r.t. the input norm, i.e., $\|\phi(\mathbf{x})-\phi(\mathbf{x}')\|_2 \leq L\|\mathbf{x}-\mathbf{x}'\|$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. Suppose input $\mathbf{x}$ satisfies Theorem 1 (i.e., low prediction entropy enforces large feature norm $\|\phi(\mathbf{x})\|_2$). Under conditions where correct prediction in Definition 1 holds, increasing the vision feature norm $\|\phi(\mathbf{x})\|_2$ expands the margin $\gamma(\mathbf{x})$. If $\gamma(\mathbf{x})>0$, the prediction for $(\mathbf{x}+\boldsymbol{\delta})$ cannot be altered away from the ground-truth label y for any perturbation $\boldsymbol{\delta} \in \mathcal{X}$ with:*

$$\|\delta\|_\infty \leq \|\delta\|_2 < \frac{\gamma(\mathbf{x})}{L\, \max_{c \neq y} \|\boldsymbol{\psi}_c - \boldsymbol{\psi}_y\|_2}. \tag{17}$$

Proof. We consider the VLM $\left(\phi(\cdot), \{\boldsymbol{\psi}_c\}_{c=1}^{C}\right)$, where $\boldsymbol{\psi}_c$ are ℓ_2-normalized, and L is the Lipschitz constant of ϕ w.r.t. the input norm:

$$\|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|_2 \;\leq\; L\,\|\mathbf{x} - \mathbf{x}'\|_2 \quad \text{for all} \quad \mathbf{x}, \mathbf{x}' \in \mathcal{X}.$$

Such Lipschitz continuity commonly holds under standard smoothness or regularity assumptions or can be enforced by spectral normalization in neural network layers, or locally by aligning adversarial samples with their clean and/or inverse adversarial counterparts at the feature level (we indeed have such an alignment realized by cross-entropy in our loss).

For an input example $\mathbf{x}$ with ground-truth category $y \in \{1, \dots, C\}$, define the *classification margin*:

$$\gamma(y|\mathbf{x}) = \boldsymbol{\psi}_y^\top \phi(\mathbf{x}) - \max_{c \neq y} \boldsymbol{\psi}_c^\top \phi(\mathbf{x}).$$

If $\gamma(y|\mathbf{x})>0$, then $\mathbf{x}$ is correctly classified by the ground-truth label y.

Part (i): Large feature norm implies a larger classification margin. By Theorem 1, low prediction entropy forces $\|\phi(\mathbf{x})\|_2$ to be sufficiently large, unless text embeddings $\boldsymbol{\psi}_c$ become unbounded. Note that the textual embeddings (prompts) are fixed and safeguarded, as they typically reside within multimodal systems and are thus not subjected to manipulations. Under the *mild alignment* assumption that $\phi(\mathbf{x})$ is not close-to-orthogonal to the ground-truth textual embedding $\boldsymbol{\psi}_y$ of class y, we examine the logit gap:

$$\boldsymbol{\psi}_y^\top \phi(\mathbf{x}) - \boldsymbol{\psi}_c^\top \phi(\mathbf{x}) = (\boldsymbol{\psi}_y - \boldsymbol{\psi}_c)^\top \phi(\mathbf{x}) = \|\boldsymbol{\psi}_y - \boldsymbol{\psi}_c\|_2 \|\phi(\mathbf{x})\|_2 \cdot \cos\left(\theta_{y,c}\right),$$

where $\theta_{y,c}$ is the angle between vectors $(\boldsymbol{\psi}_y - \boldsymbol{\psi}_c)$ and $\phi(\mathbf{x})$. Thus, increasing the vision feature norm $\|\phi(\mathbf{x})\|_2$ amplifies each logit gap provided $\cos\left(\theta_{y,c}\right)$ remains bounded away from zero. Hence, the overall classification margin:

$$\gamma(\mathbf{x}) = \min_{c \neq y} \left[(\boldsymbol{\psi}_y - \boldsymbol{\psi}_c)^\top \phi(\mathbf{x})\right]$$

also grows as $\|\phi(\mathbf{x})\|_2$ increases, leading to a wider separation from the nearest decision boundary.

Part (ii): Margin implies robustness. A standard argument in margin-based classification shows that if $\gamma(\mathbf{x})$ is large, then no small input perturbations flip the predicted label. Formally, if $\gamma(\mathbf{x}) > 0$, to flip the label from ground-truth y to some $c \neq y$, one needs:

$$\boldsymbol{\psi}_c^\top \phi(\mathbf{x}+\boldsymbol{\delta}) \geq \boldsymbol{\psi}_y^\top \phi(\mathbf{x}+\boldsymbol{\delta}) \quad \Longleftrightarrow \quad (\boldsymbol{\psi}_c - \boldsymbol{\psi}_y)^\top \left[\phi(\mathbf{x}+\boldsymbol{\delta}) - \phi(\mathbf{x})\right] \geq -\gamma(\mathbf{x}).$$

According to the Cauchy-Schwarz inequality, one can obtain:

$$\left|(\boldsymbol{\psi}_c - \boldsymbol{\psi}_y)^\top [\phi(\mathbf{x}+\boldsymbol{\delta}) - \phi(\mathbf{x})]\right| \leq \|\boldsymbol{\psi}_c - \boldsymbol{\psi}_y\|_2 \|\phi(\mathbf{x}+\boldsymbol{\delta}) - \phi(\mathbf{x})\|_2.$$

Since ϕ is L-Lipschitz, *i.e.*, $\|\phi(\mathbf{x}+\boldsymbol{\delta})-\phi(\mathbf{x})\|_2 \leq L\|\boldsymbol{\delta}\|_2$. Therefore, we have:

$$(\boldsymbol{\psi}_c-\boldsymbol{\psi}_y)^\top[\phi(\mathbf{x}+\boldsymbol{\delta})-\phi(\mathbf{x})] \geq -\|\boldsymbol{\psi}_c-\boldsymbol{\psi}_y\|_2 L\,\|\boldsymbol{\delta}\|_2.$$

Hence, if $\|\boldsymbol{\delta}\|_\infty \leq \|\boldsymbol{\delta}\|_2 < \frac{\gamma(\mathbf{x})}{L\,\|\boldsymbol{\psi}_c-\boldsymbol{\psi}_y\|_2}$, the above inequality *cannot* hold, preventing a label flip. Taking the worst-case class $c \neq y$ via $\max_{c\neq y}\|\boldsymbol{\psi}_c-\boldsymbol{\psi}_y\|_2$ establishes the radius:

$$\frac{\gamma(\mathbf{x})}{L\,\max_{c\neq y}\|\boldsymbol{\psi}_c-\boldsymbol{\psi}_y\|_2}.$$

Thus, no perturbation $\|\boldsymbol{\delta}\|_2$ below that threshold can alter the predicted label from y.

□

C.2 Proof of Theorem 3

Theorem 5 (Theorem 3 from the main text)**.** *Let $\mathbf{p}_\mathcal{T}(\mathbf{x})$ and $\mathbf{p}_\mathcal{S}(\mathbf{x})$ denote the softmax predictions over C categories w.r.t. the* teacher *and* student *VLM during weak-to-strong generalization, respectively. Suppose $\hat{\mathbf{x}} = \mathbf{x}+\boldsymbol{\delta}$ is an* adversarial example *for student, while $\check{\mathbf{x}} = \mathbf{x}+\check{\boldsymbol{\delta}}$ is an* inverse *adversarial example for the teacher with high confidence on the ground-truth class. Then the following cross-entropy inequality holds:*

$$\mathcal{L}_{CE}\big(\mathbf{p}_\mathcal{S}(\hat{\mathbf{x}}),\mathbf{p}_\mathcal{T}(\check{\mathbf{x}})\big) \geq \mathcal{L}_{CE}\big(\mathbf{p}_\mathcal{S}(\mathbf{x}),\mathbf{p}_\mathcal{T}(\check{\mathbf{x}})\big), \tag{18}$$

Proof. Recall that $\mathcal{L}_{\mathrm{CE}}(\mathbf{q},\mathbf{p}) = -\sum_{i=1}^{C} p_i \log(q_i)$ is the cross-entropy of $\mathbf{q}$ w.r.t. the target distribution $\mathbf{p}$. Let teacher's prediction on inverse adversaries $\mathbf{p}_\mathcal{T}(\check{\mathbf{x}}) = [\check{p}_1,\check{p}_2,\cdots,\check{p}_C]$, student's prediction on adversaries $\mathbf{p}_\mathcal{S}(\hat{\mathbf{x}}) = [\hat{q}_1,\hat{q}_2,\cdots,\hat{q}_C]$, and student's prediction on clean samples $\mathbf{p}_\mathcal{S}(\mathbf{x}) = [q_1,q_2,\cdots,q_C]$. The cross-entropy difference in Eq. (18) can be written as:

$$\mathcal{L}_{\mathrm{CE}}\big(\hat{\mathbf{q}},\mathbf{p}_\mathcal{T}(\check{\mathbf{x}})\big)-\mathcal{L}_{\mathrm{CE}}\big(\mathbf{q},\mathbf{p}_\mathcal{T}(\check{\mathbf{x}})\big) = -\sum_{i=1}^{C}\check{p}_i\log(\hat{q}_i)+\sum_{i=1}^{C}\check{p}_i\log(q_i) = \sum_{i=1}^{C}\check{p}_i\log\left(\tfrac{q_i}{\hat{q}_i}\right), \tag{19}$$

where $\hat{\mathbf{q}} := \mathbf{p}_\mathcal{S}(\hat{\mathbf{x}})$ and $\mathbf{q} := \mathbf{p}_\mathcal{S}(\mathbf{x})$. The inequality in Eq. (18) is thus equivalent to:

$$\sum_{i=1}^{C}\check{p}_i\log\left(\tfrac{q_i}{\hat{q}_i}\right) \geq 0.$$

By construction, $\check{\mathbf{x}}$ is an *inverse* adversarial example for the teacher with high confidence in category 1. We assume the high-confidence prediction w.r.t. the inverse adversary $\check{p}_1 \gg \check{p}_i$ for index $i>1$. In addition, let $\hat{\mathbf{x}}$ be an adversarial example against the student model, which means that its prediction $\hat{\mathbf{q}}$ reduces the probability of the ground-truth class 1 by some positive amount $\kappa_1 > 0$ and increases the probability of each wrong category $i>1$ by $\kappa_i > 0$. Thus, we obtain:

$$\hat{q}_1 = q_1-\kappa_1, \quad \hat{q}_i = q_i+\kappa_i, \quad i = 2,\ldots,C.$$

This adversarial shift captures the typical adversarial behavior moving mass *away* from the ground-truth class 1 to the remaining classes.

We thus examine the sum:

$$\sum_{i=1}^{C}\check{p}_i\log\left(\tfrac{q_i}{\hat{q}_i}\right) = \check{p}_1\log\left(\tfrac{q_1}{q_1-\kappa_1}\right)+\sum_{i=2}^{C}\check{p}_i\log\left(\tfrac{q_i}{q_i+\kappa_i}\right).$$

Define:

$$\mathcal{G}(\kappa_1,\ldots,\kappa_C) = \check{p}_1\log\left(\tfrac{q_1}{q_1-\kappa_1}\right)+\sum_{i=2}^{C}\check{p}_i\log\left(\tfrac{q_i}{q_i+\kappa_i}\right).$$

We show that for $\kappa_i > 0$, $\mathcal{G}(\kappa) \geq 0$. Its derivative w.r.t. κ_1 is:

$$\frac{\partial\mathcal{G}}{\partial\kappa_1} = \frac{\check{p}_1}{q_1-\kappa_1}, \quad \text{and for } i>1, \quad \frac{\partial\mathcal{G}}{\partial\kappa_i} = -\frac{\check{p}_i}{q_i+\kappa_i}.$$

Hence, the Hessian is diagonal with entries:

$$\frac{\partial^2 \mathcal{G}}{\partial \kappa_1^2} = \frac{\check{p}_1}{(q_1 - \kappa_1)^2} > 0, \quad \frac{\partial^2 \mathcal{G}}{\partial \kappa_i^2} = \frac{\check{p}_i}{(q_i + \kappa_i)^2} > 0, \quad (i \geq 2).$$

Thus, $\mathcal{G}$ is *strictly convex* in the vector κ. By the first-order property of convex functions, for any κ, κ^*, we get:

$$\mathcal{G}(\kappa) \geq \mathcal{G}(\kappa^*) + \nabla \mathcal{G}(\kappa^*)^\top (\kappa - \kappa^*).$$

Evaluating at $\kappa^* = \mathbf{0}$. Set $\kappa^* = (0, 0, \ldots, 0)$; then we get:

$$\mathcal{G}(\mathbf{0}) = \check{p}_1 \log\left(\tfrac{q_1}{q_1}\right) + \sum_{i=2}^{C} \check{p}_i \log\left(\tfrac{q_i}{q_i}\right) = 0.$$

Moreover,

$$\nabla \mathcal{G}(\mathbf{0}) = \left(\frac{\check{p}_1}{q_1}, -\frac{\check{p}_2}{q_2}, \ldots, -\frac{\check{p}_C}{q_C}\right).$$

Hence, for any $\kappa = (\kappa_1, \ldots, \kappa_C)$,

$$\mathcal{G}(\kappa) \geq 0 + \sum_{i=1}^{C} \left(\frac{\partial \mathcal{G}}{\partial \kappa_i}\Big|_{\kappa^*=\mathbf{0}}\right) \kappa_i = \kappa_1 \left[\tfrac{\check{p}_1}{q_1}\right] - \sum_{i=2}^{C} \kappa_i \left[\tfrac{\check{p}_i}{q_i}\right].$$

Ensuring non-negativity of Eq. (19). Since $\check{p}_1 \gg \check{p}_i$ for $i > 1$, we may assume $\frac{\check{p}_1}{q_1} \geq 1$ and $\frac{\check{p}_i}{q_i} \leq 1$ for $i > 1$. (This aligns with the intuition that the teacher distribution over the inverse adversarial example $\mathbf{p}_{\mathcal{T}}(\check{\mathbf{x}})$ is heavily concentrated on the ground-truth class 1, whereas $\mathbf{p}_{\mathcal{S}}(\mathbf{x})$ is less confident.) Thus, we get the following

$$\kappa_1 \tfrac{\check{p}_1}{q_1} \geq \kappa_1, \quad \kappa_i \tfrac{\check{p}_i}{q_i} \leq \kappa_i, \quad i = 2, \ldots, C.$$

Therefore,

$$\mathcal{G}(\kappa) \geq \kappa_1 - \sum_{i=2}^{C} \kappa_i = 0,$$

implying $\sum_{i=1}^{C} \check{p}_i \log\left(\frac{q_i}{\hat{q}_i}\right) \geq 0$. Revisiting Eq. (19) shows that the cross-entropy difference is non-negative. Hence, we obtain:

$$\mathcal{L}_{\text{CE}}\big(\hat{\mathbf{q}}, \mathbf{p}_{\mathcal{T}}(\check{\mathbf{x}})\big) \geq \mathcal{L}_{\text{CE}}\big(\mathbf{q}, \mathbf{p}_{\mathcal{T}}(\check{\mathbf{x}})\big) \quad \Longleftrightarrow \quad \mathcal{L}_{\text{CE}}\big(\mathbf{p}_{\mathcal{S}}(\hat{\mathbf{x}}), \mathbf{p}_{\mathcal{T}}(\check{\mathbf{x}})\big) \geq \mathcal{L}_{\text{CE}}\big(\mathbf{p}_{\mathcal{S}}(\mathbf{x}), \mathbf{p}_{\mathcal{T}}(\check{\mathbf{x}})\big),$$

which completes the proof of Eq. (18).

□

D Objective Functions for Adversary and Inverse Adversary Generation

In this section, we provide several alternatives for both adversary and inverse adversary generation in an unsupervised manner during our proposed Adv-W2S learning.

D.1 Adversary Generation in Weak-to-strong Generalization

In addition to generating adversarial examples by maximizing teacher-student prediction distance in Eq. (12), we here also consider two adversary generation schemes solely based on the student VLM: (i) feature-level deviation between clean and adversarial examples, and (ii) prediction-level deviation between clean and adversarial examples, as follows:

Student feature deviation ($\boldsymbol{\delta}_{\text{S-Feat}}$ as adversarial perturbations):

$$\boldsymbol{\delta}_{\text{S-Feat}} = \arg \max_{\|\boldsymbol{\delta}_{\text{S-Feat}}\|_\infty \leq \epsilon} \big\| f_{\mathcal{S}}(\mathbf{x} + \boldsymbol{\delta}_{\text{S-Feat}}), f_{\mathcal{S}}(\mathbf{x}) \big\|_2^2, \tag{20}$$

where $f_{\mathcal{S}}(\cdot)$ denotes the image encoder of the student CLIP model during fine-tuning.

Student prediction deviation ($\boldsymbol{\delta}_{\text{S-Pred}}$ as adversarial perturbations):

$$\boldsymbol{\delta}_{\text{S-Pred}} = \arg \max_{\|\boldsymbol{\delta}_{\text{S-Pred}}\|_\infty \leq \epsilon} \mathcal{L}_{\text{CE}}\big(\mathbf{p}_{\mathcal{S}}(\mathbf{x} + \boldsymbol{\delta}_{\text{S-Pred}}), \mathbf{p}_{\mathcal{S}}(\mathbf{x})\big). \tag{21}$$

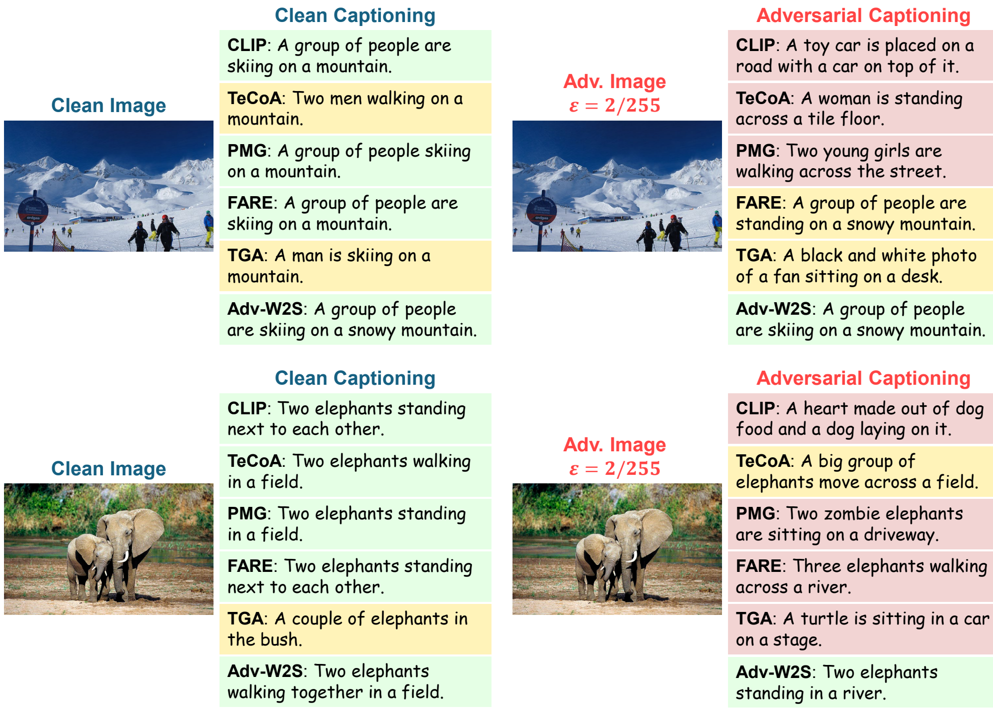


Figure 1: Visual examples (clean and adversarial samples) for **image captioning** when using the LLaVA 1.5 framework with the robust vision encoder of diverse adversarial VLM learning methods.

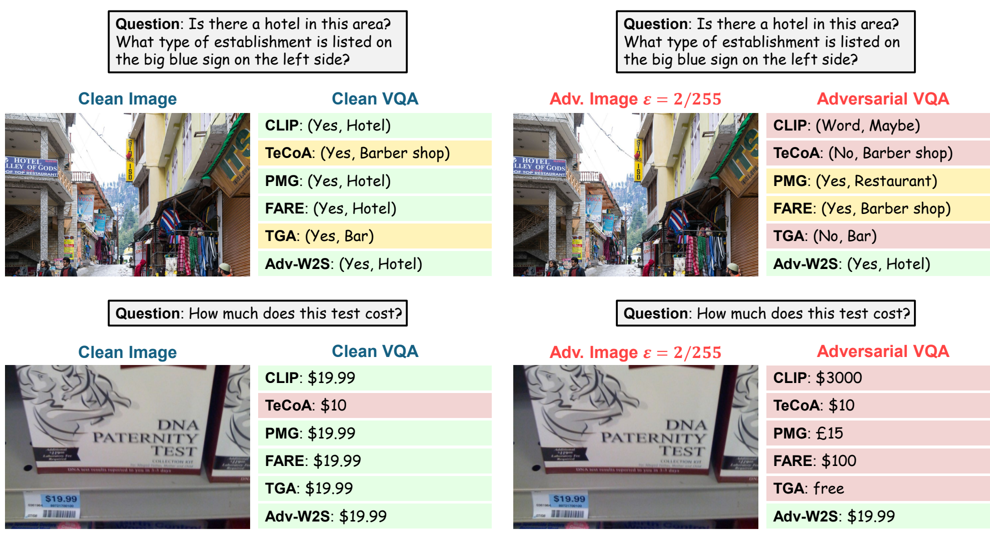


Figure 2: Visualizations (clean and adversarial samples) for **visual question answering** based on LLaVA 1.5 with the robust vision encoder of diverse adversarial VLM learning methods.

D.2 Inverse Adversary Generation in Weak-to-strong Generalization

Recall that in the main text, we focus on inverse adversary generation based on the cross-entropy loss with reference to the one-hot vector based on the teacher's prediction $\mathcal{M}(\mathbf{p}_{\mathcal{S}}(\mathbf{x}))$ (see Eq. (12)). We further consider two other types of inverse adversary generation in an unsupervised scheme: (i) pseudo-margin minimization, and (ii) prediction entropy minimization, as follows:

Teacher pseudo-margin minimization ($\check{\boldsymbol{\delta}}_{\text{T-M}}$ as adversarial perturbations):

$$\check{\boldsymbol{\delta}}_{\text{T-M}} = \arg \min_{\|\check{\boldsymbol{\delta}}_{\text{T-M}}\|_\infty \leq \epsilon} \big[\max_j \mathbf{p}_{\mathcal{T}}^j(\mathbf{x} + \check{\boldsymbol{\delta}}_{\text{T-M}}) - \max_{k \neq j^*} \mathbf{p}_{\mathcal{T}}^k(\mathbf{x} + \check{\boldsymbol{\delta}}_{\text{T-M}}) \big], \tag{22}$$

where $j^* = \arg \max_j \mathbf{p}_{\mathcal{T}}^j(\mathbf{x} + \check{\boldsymbol{\delta}})$ denotes the index of the maximum prediction.

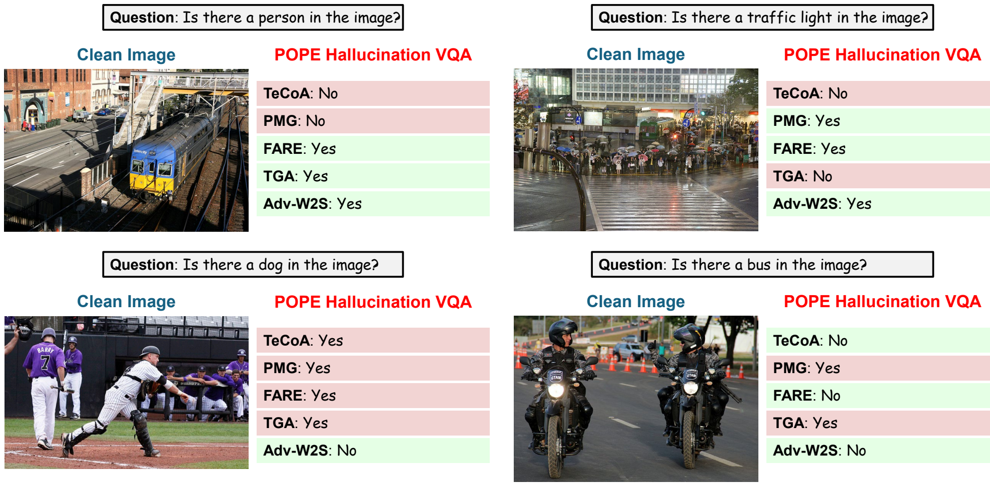


Figure 3: Visual examples for **POPE hallucination evaluations** when using the LLaVA 1.5 framework with the robust vision encoder of diverse adversarial VLM learning methods.

Teacher prediction entropy minimization ($\check{\boldsymbol{\delta}}_{\text{T-PredEP}}$ as adversarial perturbations):

$$\check{\boldsymbol{\delta}}_{\text{T-PredEP}} = \arg \min_{\left\|\check{\boldsymbol{\delta}}_{\text{T-PredEP}}\right\|_{\infty} \leq \epsilon} \left[\max \mathsf{H}_{\mathrm{T}}(\mathbf{x} + \check{\boldsymbol{\delta}}_{\text{T-PredEP}}) \right], \tag{23}$$

where $\mathsf{H}_{\mathrm{T}}(\cdot)$ denotes the prediction-level entropy of the teacher VLM.

E Auxiliary Ground-truth Supervision

Recall that we primarily concentrate on robust weak-to-strong generalization in an unsupervised scheme (see our Adv-W2S method in Eq. (13)). Despite its improved zero-shot performance as shown in Table 2, its corresponding in-distribution performance on the fine-tuned dataset is still lower than standard supervised adversarial fine-tuning (*e.g.*, TeCoA [55]) at the same VLM backbone. To mitigate such an in-distribution performance drop, we investigate an auxiliary branch of ground-truth supervision for our standard Adv-W2S method (Eq. (13)):

$$\mathcal{L}_{\text{Adv-AuxConf}}^{\textbf{Sup}\text{-W-Inv}} = \mathcal{L}_{\text{Adv-AuxConf}}^{\text{W-Inv}} + \lambda_{\text{Sup}} \cdot \mathcal{L}_{\text{CE}}\big(\mathbf{p}_{\mathcal{S}}(\mathbf{x} + \boldsymbol{\delta}), \mathbf{y}(c)\big), \tag{24}$$

where $\lambda_{\text{Sup}} = 0.5$ denotes the weighting factor for the auxiliary supervised loss. The corresponding hyper-parameter analysis is shown in Figures 5c & 5d. The trade-off effect of such an auxiliary supervision is demonstrated in Table 13.

F Visualizations of Image Captioning and Visual Question Answering

We here provide visualizations of diverse adversarial learning approaches (including our proposed Adv-W2S method) in the context of image captioning and visual question answering.

Image captioning. As shown in Figure 1, we can observe that our Adv-W2S method facilitates an adversarial invariance when confronted with unforeseen adversaries, generating correct and high-quality captioning results for both clean samples and their adversarial counterparts. While most baselines exhibit severe semantic drift or hallucinated content under adversarial perturbations, such as describing toy cars, turtles, or zombie elephants, our method, Adv-W2S, maintains robust and semantically consistent captions across both scenarios. These results highlight the superior robustness of Adv-W2S in preserving grounded visual-language alignment under distributional shift.

Visual question answering. Figure 2 illustrates the robustness of the LLaVA framework using the robust vision encoder w.r.t. diverse adversarial VLM learning approaches under clean and adversarial conditions. For clean images, most methods provide consistent and correct answers. However, under adversarial perturbations ($\epsilon = 2/255$), other comparative approaches exhibit severe degradation, producing hallucinated or even incorrect answers, e.g., listing the price as "\$3000" or

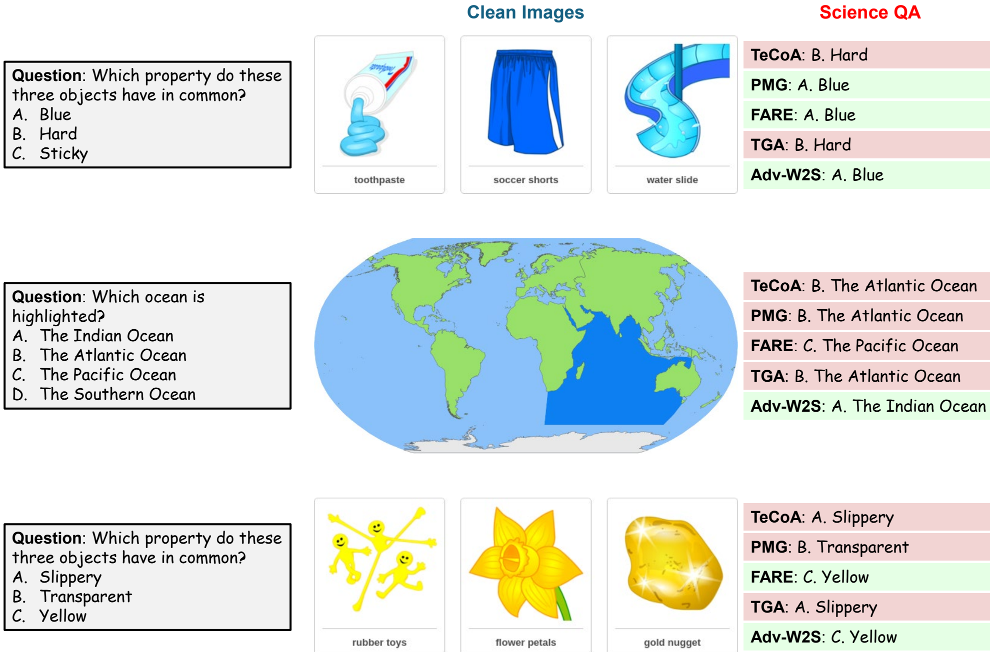
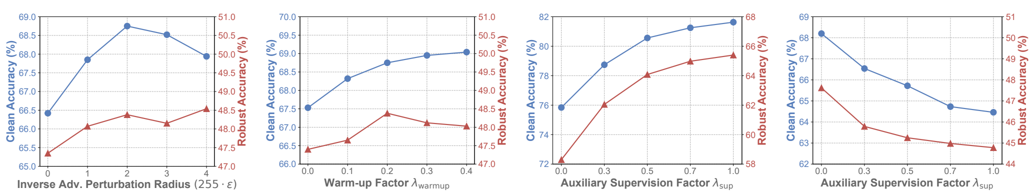


Figure 4: Visual examples for **science question answering** when using the LLaVA 1.5 framework with the robust vision encoder of diverse adversarial VLM learning methods.



(a) $\check{\epsilon}$ (Avg.) (b) λ_{warmup} (Avg.) (c) λ_{sup} (In-Distribution) (d) λ_{sup} (OOD)

Figure 5: Hyper-parameter (inverse perturbation radius $\check{\epsilon}$, warm-up factor λ_{warmup}, and auxiliary supervision factor λ_{sup}) sensitivity of Adv-W2S. (a) Varying $\check{\epsilon}$ averaged across 14 datasets. (b) Different λ_{warmup} setups averaged across 14 datasets. (c) Sensitivity to λ_{sup} on ImageNet (in-distribution). (d) Sensitivity to λ_{sup} averaged over the remaining 13 Out-Of-Distribution (OOD) datasets.

misdentifying a hotel as a "barber shop". In contrast, our Adv-W2S method consistently preserves correct and semantically grounded answers, demonstrating strong zero-shot robustness against adversarial perturbations in real-world VQA scenarios.

POPE object hallucination. The visualizations of the POPE benchmark [44] in Figure 3 illustrate how our Adv-W2S method suppresses object-level hallucinations in the adversarial evaluation scenario. Note that previous adversarial VLM learning works tend to over-rely on language priors: they answer "Yes" whenever the question mentions a common object (person and traffic light) and "No" when it mentions an unlikely one (dog on a baseball field), regardless of the actual pixels. However, our proposed Adv-W2S is the only method that consistently avoids hallucinating objects that are absent or missing objects that are present.

Science question answering. Figure 4 focuses on multi-step reasoning under the ScienceQA protocol [52]. Although all methods observe the same inputs, other adversarial VLM learning approaches collapse to the most linguistically co-occurring option (e.g., Atlantic Ocean for a map highlighting the Indian Ocean). However, our Adv-W2S maintains the correct answer even when it contradicts the majority of peer models. The result supports our claim that adversarial weak-to-strong generalization not only reduces single-token hallucinations but also strengthens higher-order inference pipelines.

G Hyper-parameter Sensitivity Analyses.

A key practical advantage of our Adv-W2S method has only a few mild-impact hyper-parameters apart from the standard parameter optimizer settings, Adv-W2S introduces only (i) the inverse adversarial perturbation $\check{\epsilon}$ that controls adversary strength in Eq. (12), (ii) the warm-up factor λ_{warmup} that anneals the adversarial loss in the early epochs, and (iii) the auxiliary supervision factor λ_{sup} in Eq. (24). Notably, no manual loss-balancing terms are required, as all three scalars modulate self-contained components and therefore have bounded influence.

We adopt the same inverse adversarial perturbation radius $\check{\epsilon}$ configuration as the standard adversary setup for consistency. We can also observe that increasing $\check{\varepsilon}$ mildly strengthens both clean and robust accuracy up to $\check{\varepsilon} = 2$, beyond which returns saturate. A minor weighting for the adversarial warm-up facilitates improving natural performance and adversarial robustness. The analyses in Figures 5c & 5d further justify the underlying trade-off between in-distribution robustness and out-of-distribution robustness. We therefore choose a balanced value of $\lambda_{\text{sup}} = 0.5$ in our standard setting, which gives near-optimal ImageNet accuracy while limiting OOD performance degradation.

H Broader Impact and Limitations.

H.1 Broader Impact

The advancement of VLMs has significantly revolutionized zero-shot learning across multimodal tasks. However, their real-world deployment raises serious concerns regarding their inherent vulnerability to adversarial attacks, which can severely compromise reliability and trustworthiness in high-stakes applications. Our proposed adversarial weak-to-strong generalization (Adv-W2S) framework directly addresses this concern by improving the zero-shot adversarial robustness of VLMs across diverse multimodal tasks in an unsupervised scheme.

Our method also aligns with the emerging objective of superalignment, which arises from the widening gap between model capabilities and the scalability of human supervision [41]. As models increasingly exceed human performance, reliance on weak teacher models to guide stronger student models becomes a scalable alternative (analogy of superalignment). Our work advances this paradigm by showing that adversarially guided supervision, rather than solely natural signals, can more effectively elicit robustness in vision-language settings.

The improvement carries several broader impacts:

- **Research advancement.** Our theoretical and empirical analyses of entropy-guided uncertainty re-weighting and inverse adversarial examples provide more insights into robust knowledge superalignment in cross-modal settings, potentially inspiring future work in robust cross-modal learning and robust AI with foundation models.
- **AI safety.** By strengthening the robustness of foundational models against unforeseen adversarial examples, our method lays the groundwork for more secure and trustworthy AI systems applicable to diverse vision recognition and understanding tasks.
- **Public trust.** As AI technologies become embedded in everyday decision-making, improving robustness under distribution shift and adversarial scenarios contributes to building public confidence in their safe and responsible deployment.

H.2 Limitations

While our method significantly enhances adversarial robustness, it presents several limitations, yet we also try to address/mitigate these limitations in our paper:

- **In-distribution performance trade-off.** The unsupervised nature of our Adv-W2S can lead to a slight degradation in in-distribution (e.g., the ImageNet dataset for fine-tuning) performance. However, we show that this can be mitigated by incorporating a supervised loss term into our optimization, allowing the model to balance zero-shot performance and in-distribution performance more effectively (see the last analysis in Section 4.3).
- **Dependence on teacher reliability.** The effectiveness of weak-to-strong generalization inherently depends on the quality of the weak teacher's predictions. When the teacher exhibits high uncer-

tainty or spurious behavior, it may misguide the student. This is a common challenge to all the weak-to-strong generalization approaches. To somehow mitigate this, our proposed framework introduces entropy-guided re-weighting and inverse adversarial refinement to reduce reliance on low-confidence supervision (see Sections 3.3 and 3.4).

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly convey the key contributions and scope of the paper with a clear motivation for robust weak-to-strong generalization. The claims are consistently supported by theoretical analyses and extensive experimental results.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The limitations are discussed in Appendix H.2.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: The paper provides complete theoretical results, with all assumptions clearly stated alongside each theorem. Full formal proofs are included in the appendix, and intuitive sketches are offered in the main text to aid understanding.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Detailed experimental setups are provided in Appendix B. For each experiment/anaylsis, we also provide its corresponding configuration for clarity.

   Guidelines:

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

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

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

   Answer: [No]

   Justification: Detailed experimental setups for reproducing our results are provided in Appendix B. All the datasets used in the paper are publicly available.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Full details of training and evaluations are provided in Appendix B.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We repeat each experiment five times under different random seeds and report the mean performance. Consistent trends across runs indicate the robustness of our findings.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The required computing resources in our setting are provided in Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We have thoroughly checked the NeurIPS Code of Ethics for our research.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Broader positive impacts are discussed in Appendix H.1.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All the external resources we used are properly mentioned and credited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

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

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

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

    Answer: [NA]

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

    Guidelines:

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

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

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

    Answer: [NA]

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

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [NA]

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

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.