

# Confound from All Sides, Distill with Resilience: Multi-Objective Adversarial Paths to Zero-Shot Robustness

Junhao Dong<sup>1,2</sup>, Jiao Liu<sup>1</sup>, Xinghua Qu<sup>3</sup>, and Yew-Soon Ong<sup>1,2\*</sup>

<sup>1</sup>Nanyang Technological University, <sup>2</sup>CFAR, IHPC, A\*STAR, <sup>3</sup>Bytedance  
{junhao003, jiao.liu, asysong}@ntu.edu.sg, xinghua.qu@bytedance.com

## Abstract

*Adversarially robust knowledge distillation transfers the robustness of a large-scale teacher model to a lightweight student while preserving natural performance. However, foundation Vision-Language Models (VLMs) also demand the transfer of zero-shot inference capabilities. We find that standard robust distillation using untargeted adversarial examples fails to transfer out-of-distribution (zero-shot) robustness, as these adversaries primarily push inputs away from their original distribution, exploring a limited portion of the teacher’s decision space and missing more diverse failure modes. A natural solution is to generate multiple targeted adversaries that traverse diverse paths across decision boundaries. Thus, these adversaries probe a broader region of the teacher’s decision surface. However, naive targeted adversary optimization often converges to local optima within a single category’s decision region, limiting the diversity. To address this, we propose a Multi-Objective Optimization (MOO)-based adversarial distillation framework that transfers robustness from large VLMs to lightweight ones by exploiting adversaries with two main objectives: misclassification and category-level adversarial diversity. Theoretically, we show that optimizing for diversity mitigates adversarial collapse into local optima, ensuring adversaries span multiple decision regions and capture the teacher’s generalizable robust features. Extensive experiments demonstrate the superiority of our method over state-of-the-art adversarial learning across diverse scenarios.*

## 1. Introduction

Vision-Language Models (VLMs) have demonstrated remarkable success across diverse domains [8, 30, 42]. However, extensive research has revealed their inherent susceptibility to adversarial perturbations [41]—subtle input modifications that are imperceptible to the human eye yet can lead to predictions with high confidence [32, 50]. Due to

their multimodal nature and increased architectural complexity, VLMs are particularly susceptible to adversarial attacks. Such vulnerabilities raise serious concerns about the reliability of VLMs in security-critical scenarios [10, 20].

To counter adversarial threats, recent works focused on improving VLM robustness via adversarial fine-tuning [21, 34, 43] with CLIP [37]. Despite its effectiveness, according to scaling laws, (robust) performance is inherently tied to model size [17, 18, 29], making large VLMs more resilient than their lightweight ones. This underlying dependence on scale presents a challenge for deploying VLMs. Adversarial distillation [15, 22] emerges as a promising alternative to address this, yet rare efforts have been dedicated to VLMs. Similar to adversarial fine-tuning, adversarial distillation relies on untargeted adversaries. However, these adversaries only push inputs away from their original distribution, limiting exploration of the teacher’s decision space and missing diverse failure modes. The restricted perturbation space hinders robustness generalization, particularly in out-of-distribution (zero-shot) tasks.

To support this claim, we analyze robustness transfer using untargeted and targeted adversaries, evaluating their impact on in-distribution and out-of-distribution robustness (Figure 1a and 1b). Specifically, *untargeted adversaries* follow a gradient ascent path away from the original class (distribution)  $c_i$  to maximize the classification error, while *targeted adversaries* take a gradient descent trajectory toward a specific class  $c_t \neq c_i$ <sup>1</sup>. Our key observation is that untargeted adversaries primarily enhance in-distribution robustness, as they expose the worst-case decision boundaries of the training distribution. In contrast, targeted adversaries improve out-of-distribution robustness by spanning multiple class boundaries, promoting generalization beyond the training set. This trade-off arises as untargeted adversaries focus on the most vulnerable points in the training distribution, potentially leading to over-fitting in distillation, while targeted adversaries explore a more diverse decision space, aiding robustness transfer across distributions. Our *Multi-*

\*Corresponding author.

<sup>1</sup>The formal definition will be provided in the following section.

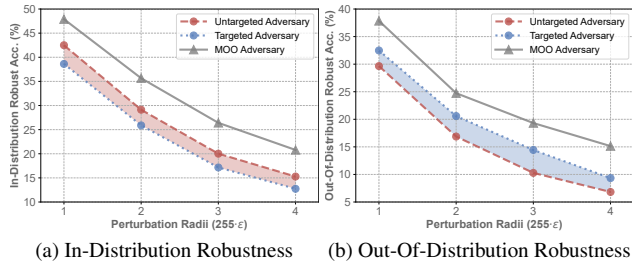


Figure 1. **Adversarial distillation using ImageNet with targeted and untargeted adversaries.** (a) In-distribution robustness on ImageNet. (b) Out-of-distribution robustness (average over SUN397, Flower102, and CIFAR-100). Robust accuracy across perturbation radii shows that untargeted adversaries facilitate transferring in-distribution robustness, while targeted adversaries improve out-of-distribution robustness, highlighting their complementary effects in our multi-objective optimization-based adversarial distillation.

*Objective Optimization (MOO)-based Adversarial Distillation (MOO-AD)* leverages these complementary effects to improve in-distribution and out-of-distribution robustness.

In our MOO-AD, we conduct robustness transfer from large-scale VLMs to lightweight ones by exploiting adversaries with two main objectives: misclassification and category-level adversarial diversity. In other words, we encourage each adversary to exhibit a distinct yet strong failure mode, ensuring mutual complementation between targeted and untargeted adversaries. This leads to a diverse set of MOO-based adversaries that comprehensively cover the teacher’s decision landscape. By simply mimicking the robust behavior (*e.g.*, predictions) of the teacher based on MOO-adversaries, we enable a more generalizable transfer of the teacher’s robust knowledge to the student model. Theoretically, we demonstrate that diversity-guided optimization mitigates adversarial collapse into local optima, ensuring adversaries span multiple decision regions and more accurately capture the teacher’s decision surface. We further show that incorporating these MOO-adversaries into distillation optimizes an upper bound of robust risk, leading to improved generalization against distribution shifts.

Extensive experiments validate MOO-AD’s superiority in zero-shot performance over state-of-the-art methods. Its generalizability is proved through extensions to medical imaging and vision-language understanding. Systematic analyses justify its effectiveness in robustness transfer.

Our core contributions can be summarized below:

- We first analyze the trade-off between in-distribution and out-of-distribution robustness in VLM adversarial distillation, linked to untargeted and targeted adversaries.
- To address this trade-off and enhance robustness transfer, we propose a novel adversarially robust knowledge distillation framework leveraging a simple yet effective multi-objective optimization to generate adversaries with diversity, capturing decision surface more comprehensively.

- We theoretically demonstrate that diversity-guided optimization mitigates adversarial collapse into local optima, ensuring adversaries uniformly span the decision surface. We further show that incorporating these adversaries into distillation minimizes an upper bound of the robust risk.
- Extensive experiments demonstrate the state-of-the-art performance of our method across various datasets, architectures, vision-language tasks, and zero-shot settings.

**Related works.** Numerous studies have demonstrated the formidable risks brought by adversaries in single-modal architectures [12, 14, 23, 48], yet recent works showed that VLMs are even more vulnerable to adversaries [5, 50, 53]. A series of remedies have been proposed to enhance the robustness [46], with adversarial fine-tuning/training [11, 13, 34, 38, 43] emerging as the most effective defense, typically with Parameter-Efficient Fine-Tuning (PEFT) [28, 52, 54]. Mao *et al.* [34] pioneered adversarial fine-tuning via image-text contrastive learning to align adversarial image features to their text counterparts. Schlarmann *et al.* [38] focused on robustness generalization across downstream tasks. Despite their effectiveness, according to scaling laws, (robust) performance is inherently tied to model size [29], resulting in a significant robustness gap between large-scale and lightweight VLMs. A natural solution is knowledge distillation [25] and its adversarial variant [22, 26] for robustness transfer, while there exist rare works in the context of VLMs. To fill this gap, we propose a novel adversarially robust VLM knowledge distillation framework based on Multi-Objective Optimization (MOO)-guided adversaries, which leverage MOO [19, 35, 55] to simultaneously balance misclassification success and adversarial diversity. (A detailed introduction about the background of MOO can be found in Appendix B.) We show that, by exploiting the complementary effects of untargeted and targeted adversaries, our method achieves a better trade-off between in-distribution and out-of-distribution robustness.

## 2. Background

CLIP [37] has demonstrated superior generalizability by integrating visual and textual representations within a shared embedding space, consisting of two key components: an image encoder, denoted as  $f_{\theta_I} : \mathcal{X} \rightarrow \mathbb{R}^d$ , and a text encoder, represented as  $f_{\theta_T} : \mathcal{T} \rightarrow \mathbb{R}^d$ , parameterized by  $\theta_I$  and  $\theta_T$ , respectively. For any given image-text pair  $(\mathbf{x}, \mathbf{t})$ , these encoders extract the feature representations from both modalities, which can lead to image-text alignment by optimizing the cosine similarity between them. Consequently, for an image input  $\mathbf{x}$ , the predicted probability of it being associated with class  $c \in \{1, \dots, C\}$  is computed as:

$$p_c(\mathbf{x}) = \frac{\exp(\cos(f_{\theta_I}(\mathbf{x}), f_{\theta_T}(\mathbf{t}_c)))}{\sum_{i=1}^C \exp(\cos(f_{\theta_I}(\mathbf{x}), f_{\theta_T}(\mathbf{t}_i)))}, \quad (1)$$

where  $\mathbf{t}_i$  follows the structure “[Context][CLASS<sub>*i*</sub>]”, such as “This is a photo of a [CLASS<sub>*c*</sub>]”, which is tokenized and transformed into an embedding via the text encoder  $f_{\theta_T}(\cdot)$ . Here  $\exp(\cdot)$  and  $\cos(\cdot)$  correspond to the exponential function and cosine similarity measure, respectively. For notational simplicity, we define the prediction (probability vector) as  $\mathbf{p} = [p_1, \dots, p_C]^\top \in \mathbb{R}_+^C$ . Hence, to enhance VLM robustness, standard adversarial fine-tuning (TeCoA [34]) optimizes the following min-max formulation of image-text alignment for a given dataset  $\mathcal{D}$ :

$$\min_{\theta_I} \mathbb{E}_{(\mathbf{x}, c) \sim \mathcal{D}} \left[ \max_{\|\delta_I\|_\infty \leq \epsilon_I} \mathcal{L}_{\text{CE}}(\mathbf{p}(\mathbf{x} + \delta_I), \mathbf{y}(c)) \right], \quad (2)$$

where  $\delta_I$  is a perturbation applied to clean sample  $\mathbf{x}$ , constrained within an  $\ell_\infty$ -norm ball of radius  $\epsilon_I$ . Untargeted adversarial examples for misclassification to random classes are represented as  $\hat{\mathbf{x}} = \mathbf{x} + \delta_I$ . The label  $c$  is encoded as a one-hot vector  $\mathbf{y}(c) = [\mathbf{1}(c = 1), \dots, \mathbf{1}(c = C)]^\top \in \{0, 1\}^C$ , where each entry indicates the presence of a specific class. Consistent with prior adversarial learning [33, 34], *untargeted adversary generation* in inner maximization optimizes the Cross-Entropy (CE) loss  $\mathcal{L}_{\text{CE}}$  through an iterative *gradient ascent* strategy at the input level:

$$\hat{\mathbf{x}}^{(i+1)} = \Pi_{\mathbb{B}(\mathbf{x}, \epsilon_I)} \left[ \hat{\mathbf{x}}^{(i)} + \alpha_I \cdot \text{sign} \left( \nabla_{\hat{\mathbf{x}}^{(i)}} \mathcal{L}_{\text{CE}}(\mathbf{p}(\hat{\mathbf{x}}^{(i)}), \mathbf{y}(c)) \right) \right], \quad (3)$$

where  $\alpha_I$  is the step size, while  $\Pi_{\mathbb{B}(\mathbf{x}, \epsilon_I)}$  denotes the projection that ensures adversaries remain within a  $\ell_\infty$ -bounded region of radius  $\epsilon_I$  centered around  $\mathbf{x}$ . The initialization of untargeted adversaries, denoted as  $\hat{\mathbf{x}}^{(0)}$ , is obtained by appending a small perturbation as  $\hat{\mathbf{x}}^{(0)} \sim \mathbf{x} + 0.001 \cdot \mathcal{N}(\mathbf{0}, \mathbf{I})$ . The final adversary is obtained as  $\hat{\mathbf{x}} = \hat{\mathbf{x}}^{(m)}$  after  $m$  steps.

Another class of adversarial examples, targeted adversarial examples, aims to induce misclassification into a specific target class  $c_t$ . Similar to untargeted adversary generation, *targeted adversaries* are constructed by minimizing the classification loss (e.g., cross-entropy) w.r.t.  $c_t$  using the following iterative *gradient descent* formulation:

$$\hat{\mathbf{x}}_{c_t}^{(i+1)} = \Pi_{\mathbb{B}(\mathbf{x}, \epsilon_I)} \left[ \hat{\mathbf{x}}_{c_t}^{(i)} - \alpha_I \cdot \text{sign} \left( \nabla_{\hat{\mathbf{x}}_{c_t}^{(i)}} \mathcal{L}_{\text{CE}}(\mathbf{p}(\hat{\mathbf{x}}_{c_t}^{(i)}), \mathbf{y}(c_t)) \right) \right], \quad (4)$$

where  $\hat{\mathbf{x}}_{c_t}^{(i)}$  denotes the targeted adversarial example at the  $i^{\text{th}}$  iteration, optimized to be misclassified as class  $c_t$ . Unlike untargeted adversaries, which perturb the input into any incorrect class, targeted adversaries follow a specific trajectory toward the decision region of a designated class  $c_t$ , enforcing a more structured misclassification pattern.

### 3. Proposed Method

We here introduce our adversarially robust knowledge distillation, MOO-AD, which transfers robustness from large-scale VLMs to lightweight ones via MOO-adversaries with diversity and disruption capabilities, as shown in Figure 2.

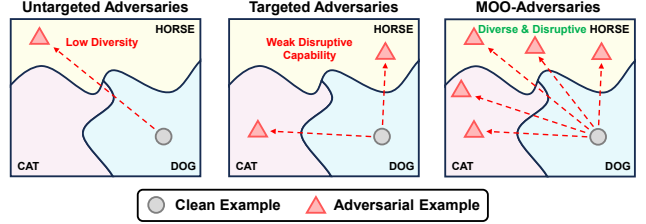


Figure 2. Illustration of untargeted, targeted, and MOO-generated adversaries. MOO-based adversaries enhance diversity and disruption, capturing a broad region of the decision surface.

**Problem definition.** Beyond standard in-distribution robustness evaluation [7], we focus on more challenging *zero-shot* robustness scenario [34]. In this setting, adversaries have unrestricted access to previously unseen datasets at inference time, whereas the defense mechanism must maintain robustness against these novel threats without prior exposure. From a practical defense perspective, textual prompts—potentially embedded within multimodal frameworks—remain unaltered at inference unless explicitly stated otherwise.

Different from adversarial fine-tuning [34], which relies solely on one-hot guidance from datasets, we focus on robustness transfer from a large robust *teacher* VLM  $[f_{\theta_T^s}(\cdot), f_{\theta_T^s}(\cdot)]$  to a lightweight *student*  $[f_{\theta_S^s}(\cdot), f_{\theta_S^s}(\cdot)]$ . The corresponding predictions of the teacher and student VLMs are indicated as  $\mathbf{p}_T(\cdot)$  and  $\mathbf{p}_S(\cdot)$ , respectively.

#### 3.1. Multi-Objective Adversary Generation

**Multi-objective optimization formulation.** In this paper, we formulate adversary generation as a MOO task:

$$\begin{aligned} \min_{\hat{\mathbf{x}}} \begin{cases} F_1(\hat{\mathbf{x}}) = \sum_{c=1}^C c \cdot [\mathbf{p}_S(\hat{\mathbf{x}})]_c, \\ F_2(\hat{\mathbf{x}}) = C - \sum_{c=1}^C c \cdot [\mathbf{p}_S(\hat{\mathbf{x}})]_c, \end{cases} \\ \text{s.t. } \mathcal{L}_{\text{CE}}(\mathbf{p}_S(\hat{\mathbf{x}}), \mathbf{y}(c)) \geq \delta_{\text{CE}}, \\ \mathcal{L}_{\text{CE}}(\mathbf{p}_T(\hat{\mathbf{x}}), \mathbf{y}(c)) \geq \delta_{\text{CE}}, \\ \hat{\mathbf{x}} \in \mathbb{B}(\mathbf{x}, \epsilon_I) \end{aligned} \quad (5)$$

where  $\delta_{\text{CE}}$  is a positive constant ensuring that the generated sample leads to incorrect predictions by the model, and  $[\mathbf{p}_S(\cdot)]_c$  represents the predicted probability assigned by the student belonging to class  $c$ . The motivation for modeling adversarial generation as a MOO problem is to enhance adversarial diversity. Intuitively, high diversity in adversaries improves decision boundary coverage, enhancing the model’s robustness generalization. Adversaries spanning different regions of the decision boundary enable more comprehensive robustness training for the student. MOO inherently seeks a diverse set of *Pareto-optimal* solutions rather than a single optimal one, making diversity mainte-

---

**Algorithm 1** Multi-Objective Adversary Generation.
 

---

```

1 # (x,y): clean image and its label
2 # (f_t, f_s): teacher and student models
3 # W: preference vector in Eq. (7)
4 # m: number of MOO-adversaries
5 # step_size: step size of adversary generation
6 # num_steps: number of adversary generation steps
7 # num_classes: number of categories
8 # epsilon: perturbation radius
9 # gamma_1, gamma_2: weighting factor in Eq.(6)
10 # CE: cross-entropy loss
11
12 X_moo = x.repeat(m,1) + random_noise
13 for _ in range(num_steps):
14     # Calculate Multi-Objective Values
15     F1 = sum(y * softmax(f_s(X_moo), dim=1), dim=1)
16     F2 = num_classes - F1
17     F = cat((F1, F2), dim=1) - gamma_1 * CE(f_t(
18         X_moo), y) + gamma_2 * CE(f_s(X_moo), y)
19     # Subproblem Decomposition
20     H = W * F
21     L_tch = H.max(dim=1) + 0.001 * H.mean(dim=1)
22     L_tch.backward()
23     # Iterative Gradient Ascent
24     eta = (-step_size) * X_moo.grad.sign()
25     eta = clamp(eta, -epsilon, epsilon)
26     X_moo = X_moo + eta
27 return X_moo

```

---

nance a well-studied topic [39, 45]. By adopting this formulation, we can leverage established diversity-preserving techniques from MOO to promote adversarial diversity.

However, it is important to note that in this work, diversity is discussed in the sample (input) space. To address this, we establish the relationship between diversity in the objective space and diversity in the sample space, assuming that  $F_1$  in Eq. (5) is  $L$ -Lipschitz continuous.

**Theorem 1.** *Assume that  $F_1$  in Eq. (5) is  $L$ -Lipschitz continuous. Let  $\mathbf{x}^a$  and  $\mathbf{x}^b$  be two Pareto optimal solutions of Eq. (5). If there exists a positive constant  $\delta_{\mathbf{F}}$  such that  $\|\mathbf{F}(\mathbf{x}^a) - \mathbf{F}(\mathbf{x}^b)\| \geq \delta_{\mathbf{F}}$ , then the distance between these solutions is bounded as  $\|\mathbf{x}^a - \mathbf{x}^b\| \geq \frac{\sqrt{2}\delta_{\mathbf{F}}}{2L}$ , where  $L$  is the Lipschitz constant, and  $\mathbf{F}(\mathbf{x}) = (F_1(\mathbf{x}), F_2(\mathbf{x}))$ .*

*Proof.* See Appendix E.1.  $\square$

**Theorem 1** shows that if two adversaries are far away from each other in the objective space of Eq. (5), they will also be far away from each other in the sample space. This implies that preserving diversity in the objective space translates to maintaining diversity among adversaries in the sample space. Consequently, this supports the feasibility of employing diversity-preserving capability in MOO algorithms to generate diverse adversaries. For better readability, we have also provided intuitive explanations in Appendix C for a better understanding of our formulated MOO problem.

**Decomposition-based MOO solver.** To circumvent the complexity of handling the constraints in Eq. (5), we instead solve its relaxed formulation, as shown below, rather

than directly addressing the original problem in Eq. (5):

$$\min_{\hat{\mathbf{x}} \in \mathbb{B}(\mathbf{x}, \epsilon_1)} \begin{cases} F'_1(\hat{\mathbf{x}}) = F_1(\hat{\mathbf{x}}) - \gamma_1 \mathcal{L}_{\text{CE}}(\mathbf{p}_S(\hat{\mathbf{x}}), \mathbf{y}(c)), \\ \quad -\gamma_2 \mathcal{L}_{\text{CE}}(\mathbf{p}_{\mathcal{T}}(\hat{\mathbf{x}}), \mathbf{y}(c)), \\ F'_2(\hat{\mathbf{x}}) = F_2(\hat{\mathbf{x}}) - \gamma_1 \mathcal{L}_{\text{CE}}(\mathbf{p}_S(\hat{\mathbf{x}}), \mathbf{y}(c)), \\ \quad -\gamma_2 \mathcal{L}_{\text{CE}}(\mathbf{p}_{\mathcal{T}}(\hat{\mathbf{x}}), \mathbf{y}(c)), \end{cases} \quad (6)$$

where  $\gamma_1$  and  $\gamma_2$  are preset weighting factors. We employ a decomposition-based approach to solve this relaxed MOO problem, obtaining a diverse set of solutions. The decomposition-based method transforms a multi-objective optimization problem into a set of single-objective subproblems through function aggregation. We thus adopt the augmented Tchebycheff scalarization [51] to achieve the function aggregation, given its compatibility with non-convex Pareto fronts. Given a preference vector  $\mathbf{w} = (w_1, w_2)$ , such that  $w_1, w_2 \geq 0$  and  $w_1 + w_2 = 1$ , the augmented Tchebycheff scalarization aggregates two objective functions into a single-objective minimization subproblem:

$$\mathcal{L}^{\text{tch}}(\hat{\mathbf{x}}|\mathbf{w}) = \max_{j \in \{1,2\}} \{w_j F'_j(\hat{\mathbf{x}})\} + \eta \sum_{j=1}^2 w_j F'_j(\hat{\mathbf{x}}), \quad (7)$$

where  $\eta$  is a tiny positive value, which we set to 0.001. The decomposition-based MOO approach provides an effective means of maintaining diversity in the objective space. This method relies on a predefined diverse set of preference vectors and the parallel solutions of the corresponding subproblems. The underlying principle can be found in [51]. Next, we present our implementation of a decomposition-based MOO solver for adversary generation.

**Multi-objective adversary generation implementation.**

Let  $\mathbf{x}$  be a sample from  $\mathcal{D}$ , then a batch of MOO adversaries  $\hat{\mathbf{X}}_{\text{MOO}} = \{\hat{\mathbf{x}}_{\text{MOO}}^k\}_{k=1}^n$  can be produced based on a preset set of preference vectors  $\mathcal{W} = \{\mathbf{w}^k\}_{k=1}^n$ <sup>2</sup>, i.e.,  $\hat{\mathbf{x}}_{\text{MOO}}^k = \arg \min_{\hat{\mathbf{x}} \in \mathbb{B}(\mathbf{x}, \epsilon_1)} \mathcal{L}^{\text{tch}}(\hat{\mathbf{x}}|\mathbf{w}^k)$ . In our implementation, we achieve this adversary generation process through an iterative gradient ascent strategy, which is:

$$\hat{\mathbf{x}}_{\text{MOO}}^{k,(i+1)} = \Pi_{\mathbb{B}(\mathbf{x}, \epsilon_1)} \left[ \hat{\mathbf{x}}_{\text{MOO}}^{k,(i)} + \alpha_1 \cdot \text{sign} \left( \nabla_{\hat{\mathbf{x}}_{\text{MOO}}^{k,(i)}} \mathcal{L}^{\text{tch}}(\hat{\mathbf{x}}_{\text{MOO}}^{k,(i)}|\mathbf{w}^k) \right) \right]. \quad (8)$$

A PyTorch-like implementation of the proposed multi-objective adversary generation is shown in **Algorithm 1**.

### 3.2. Adv. Distillation Guided by MOO-Adversaries

Unlike conventional knowledge distillation [25], which relies solely on clean examples and thus transfers non-robust features, robust knowledge distillation further enhances resilience by incorporating adversarial samples, capturing robust features. In this paper, we go beyond single adversarial points (either untargeted or targeted adversaries) to MOO-adversaries that comprehensively cover the robust decision

<sup>2</sup>We focus on the Riesz  $s$ -energy method [2] to produce  $\mathcal{W}$  in this paper.

surface of the teacher VLM, leading to a better representation of its robust knowledge. Thus, we define the Multi-Objective Prediction Alignment (MOPA) via an instance-wise teacher-student prediction alignment not only for clean samples  $\mathbf{x}$  but also for MOO-adversaries  $\hat{\mathbf{X}}_{\text{MOO}}$  to transfer both standard generalization and adversarial robustness.

$$\mathcal{L}_{\text{MOPA}} = \underbrace{\mathcal{L}_{\text{KL}}(\mathbf{p}_{\mathcal{T}}(\mathbf{x}) \parallel \mathbf{p}_{\mathcal{S}}(\mathbf{x}))}_{\text{"clean distributions" alignment}} + \beta \cdot \underbrace{\frac{\sum_{i=1}^n \mathcal{L}_{\text{KL}}(\mathbf{p}_{\mathcal{T}}(\hat{\mathbf{x}}_{\text{MOO}}^i) \parallel \mathbf{p}_{\mathcal{S}}(\hat{\mathbf{x}}_{\text{MOO}}^i))}{n}}_{\text{"adversarial distributions" alignment}}, \quad (9)$$

where  $\beta \geq 0$  regulates the trade-off between clean and adversarial prediction alignment via the Kullback–Leibler (KL) divergence  $\mathcal{L}_{\text{KL}}(\cdot \parallel \cdot)$ . Notably, we align the student’s predictions on MOO-adversaries with those of the teacher, rather than clean sample predictions. This design choice is motivated by the need to maintain a consistent data view between teacher and student VLMs, thereby mitigating the inductive bias toward local invariance at the prediction level. Without this alignment, adversarial robustness transfer can lead to over-correction [40], resulting in excessive model smoothness that hinders robustness generalization.

Beyond instance-wise prediction alignment inside the MOO-adversary set, we further enhance the robust knowledge transfer by aligning the overall distributional structure between the teacher and student VLMs. To enforce higher-order statistical consistency across adversarial predictions, we introduce a simple yet effective Group-wise Distribution Matching (GDM) scheme based on the Maximum Mean Discrepancy (MMD) with the kernel metric as follows:

$$\mathcal{L}_{\text{GDM}} = \text{MMD}^2(\{\mathbf{p}_{\mathcal{T}}(\hat{\mathbf{x}}_{\text{MOO}}^i)\}_{i=1}^n, \{\mathbf{p}_{\mathcal{S}}(\hat{\mathbf{x}}_{\text{MOO}}^i)\}_{i=1}^n), \quad (10)$$

where the discrepancy  $\text{MMD}^2(\mathcal{P}, \mathcal{Q}) = \mathbb{E}_{\mathbf{z}, \mathbf{z}' \sim \mathcal{P}}[k(\mathbf{z}, \mathbf{z}')] + \mathbb{E}_{\mathbf{w}, \mathbf{w}' \sim \mathcal{Q}}[k(\mathbf{w}, \mathbf{w}')] - 2\mathbb{E}_{\mathbf{z} \sim \mathcal{P}, \mathbf{w} \sim \mathcal{Q}}[k(\mathbf{z}, \mathbf{w})]$ . Here,  $k(\mathbf{z}, \mathbf{w})$  is a kernel function, such as Gaussian RBF kernel or polynomial kernel, ensuring matching beyond first-order and second-order statistics. Aligning adversarial distributions via GDM further prevents excessive model smoothness, which can arise when the student overcompensates for adversarial perturbations by overly suppressing prediction variance. From a functional perspective, the kernel-based GDM captures global geometric relations within the adversarial prediction space, offering a more stable optimization landscape during adversarial distillation.

**Objective function.** The overall loss function of our MOO-AD integrates two main components: multi-objective prediction alignment  $\mathcal{L}_{\text{MOPA}}$  and group-wise distribution matching  $\mathcal{L}_{\text{GDM}}$ , formulated as follows:

$$\mathcal{L} = \mathcal{L}_{\text{MOPA}} + \lambda \cdot \mathcal{L}_{\text{GDM}}, \quad (11)$$

where  $\lambda > 0$  controls the relative contributions. Following prior robust VLM studies [34, 38, 43], the network parameter updating is enabled solely for the vision encoder of CLIP for stability. At inference, the distilled student model is employed directly for robustness evaluation.

### 3.3. MOO-AD Minimizes a Robust Risk Bound

The trade-off between natural accuracy and adversarial robustness is particularly pronounced in VLMs [43]. Following the robust risk formulation [49], we adopt their decomposition into natural and boundary components associated with the derived robust risk upper bound of intermediate adversarial examples [16] as our theoretical foundation.

**Key result.** We extend this framework to the MOO setting and prove that incorporating MOO-adversarial examples into training yields an improved robust risk characterization. Specifically, when the boundary-risk gap  $\kappa$  between clean data and correctly classified MOO-adversaries satisfies  $\kappa \geq \mathcal{R}_{\text{nat}}(\mathcal{D})$ , the robust risk on the augmented dataset  $\mathcal{D} \cup \mathcal{M}_{\mathbf{x}}$  is strictly tighter than that of the original dataset  $\mathcal{D}$ . We empirically confirm in Section 5.1 that this condition consistently holds across our adversarially robust knowledge distillation experiments. Complete definitions, decompositions, and proofs are provided in Appendix D.

## 4. Experiments

### 4.1. Experimental Setups

**Datasets.** Following prior works [34, 38], we assess in-distribution robustness on ImageNet [9] and zero-shot robustness on 14 datasets. More details are in Appendix A.1.

**Implementation details.** We use CLIP [37] with ViT-L/14 as the teacher and ViT-B/32 & ResNet-50/101 as the student. MOO-adversaries are generated with 10 iterations under the  $\ell_{\infty}$  radius  $\epsilon_1 = 2/255$ . The MOO weighting factors are set  $\gamma_1 = \gamma_2 = 1.0$  for balanced optimization. The loss coefficients are  $\lambda = 2.5$  and  $\beta = 4.0$ . Evaluations are conducted under adaptive attacks. Details are in Appendix A.2.

### 4.2. Main Results (Zero-Shot Classification)

**Robustness transfer from a large-scale VLM.** To evaluate both standard generalization and adversarial robustness, we present a comparative analysis of our MOO-AD framework against state-of-the-art adversarial fine-tuning methods in Table 1&2. Beyond ImageNet, we assess zero-shot generalization across 14 extra benchmarks, measuring accuracy on both original images and their adversaries crafted by Projected Gradient Descent (PGD) [33] of 20 iterations with a fixed perturbation magnitude  $\epsilon_1 = 2/255$ . According to Table 1, our MOO-AD achieves an average improvement of 2.3%, significantly closing the performance gap with the vanilla CLIP model. We further show that our MOO-AD enhances adversarial robustness by an average of 4.8%.

**Robustness transfer of different student VLMs.** Beyond ViT-B for the student VLM, we extend our evaluations to additional CLIP backbones. For our adversarial distillation, robustness is transferred from an adversarially pre-trained ViT-L VLM. Other adversarial fine-tuning methods optimize the CLIP with sole guidance from datasets. In Table

Table 1. Adversarial distillation (ViT-L  $\rightarrow$  ViT-B, ImageNet) with evaluations across 15 datasets, reporting zero-shot **clean accuracy** (%).

Architecture	Method	ImageNet	STL10	CIFAR-10	CIFAR-100	SUN397	Stanf.Cars	Food101	OxfordPet	Flower102	DTD	EuroSAT	FGVC	PCAM	Caltech101	Caltech256	Average
ViT-L (Teacher)	TeCoA [34]	73.61	98.73	92.69	73.45	68.20	64.31	84.31	92.16	72.57	49.44	36.46	29.51	51.04	89.22	88.62	70.95
	CLIP [37]	59.13	97.17	88.55	62.29	57.68	52.07	83.84	87.35	65.60	40.05	38.31	20.13	52.26	87.08	82.01	64.90
ViT-B (Student)	TeCoA [34]	54.43	91.10	72.77	41.31	44.71	22.06	39.27	75.06	38.36	29.46	22.91	10.55	42.37	77.11	70.91	48.83
	PMG-FT [43]	51.33	90.70	73.05	42.04	44.40	27.72	42.61	75.39	39.37	29.02	20.32	11.45	47.22	80.08	71.01	49.71
	FARE [38]	50.94	93.90	81.98	56.25	49.94	42.73	63.30	81.59	53.29	34.27	21.53	14.66	45.04	85.72	75.03	56.68
	MOO-AD	<b>59.28</b>	<b>95.16</b>	<b>85.13</b>	<b>59.24</b>	<b>52.35</b>	<b>43.28</b>	<b>66.01</b>	<b>82.78</b>	<b>55.63</b>	<b>35.15</b>	<b>22.07</b>	<b>14.97</b>	<b>49.81</b>	<b>86.42</b>	<b>77.14</b>	<b>58.96</b>

Table 2. Adversarial distillation (ViT-L  $\rightarrow$  ViT-B, ImageNet) with evaluations on 15 datasets, reporting zero-shot **robust accuracy** (%) against adversaries generated via 20-step PGD adversarial attack scheme [33] with the image-level **perturbation radius** of  $\epsilon_1 = 2/255$ .

Architecture	Method	ImageNet	STL10	CIFAR-10	CIFAR-100	SUN397	Stanf.Cars	Food101	OxfordPet	Flower102	DTD	EuroSAT	FGVC	PCAM	Caltech101	Caltech256	Average
ViT-L	TeCoA [34]	52.15	91.57	74.47	50.20	49.11	38.70	49.24	79.30	50.85	42.80	15.39	14.33	50.31	80.59	73.84	54.19
ViT-B	CLIP [37]	0.54	21.38	2.21	0.73	0.35	0.20	6.11	2.99	0.54	0.22	0.03	0.00	0.00	13.60	8.81	3.85
	TeCoA [34]	27.03	71.60	44.34	23.13	18.67	5.13	13.29	42.33	16.29	17.52	12.28	2.52	13.47	56.82	45.52	27.33
	PMG-FT [43]	26.20	71.91	45.72	23.41	18.87	6.89	14.62	43.14	16.72	18.02	12.59	2.73	21.39	58.25	46.21	28.44
	FARE [38]	24.57	77.28	52.91	30.37	17.85	9.66	18.20	46.26	18.46	19.53	10.24	2.82	23.36	62.95	49.59	30.94
	MOO-AD	<b>36.58</b>	<b>78.09</b>	<b>58.77</b>	<b>32.95</b>	<b>21.15</b>	<b>10.12</b>	<b>20.01</b>	<b>55.36</b>	<b>20.13</b>	<b>19.97</b>	<b>11.85</b>	<b>3.79</b>	<b>45.28</b>	<b>66.59</b>	<b>54.86</b>	<b>35.70</b>

Table 3. Comparison of MOO-AD and adversarial fine-tuning on average performance on 15 datasets across CLIP architectures.

Architecture	Method	Clean	PGD	CW	AA
ViT-B	TeCoA [34]	48.83	27.33	26.80	25.75
	PMG-FT [43]	49.71	28.44	27.63	26.98
	FARE [38]	56.68	30.94	30.26	29.30
	MOO-AD	<b>58.96</b>	<b>35.70</b>	<b>34.95</b>	<b>34.16</b>
ResNet-50	TeCoA [34]	43.25	23.20	22.53	21.74
	PMG-FT [43]	45.11	24.68	23.72	23.05
	FARE [38]	45.43	24.27	23.19	22.24
	MOO-AD	<b>47.29</b>	<b>26.33</b>	<b>25.42</b>	<b>24.58</b>
ResNet-101	TeCoA [34]	44.96	26.61	25.42	24.93
	PMG-FT [43]	46.28	27.96	26.84	26.29
	FARE [38]	48.16	22.32	21.71	21.18
	MOO-AD	<b>49.82</b>	<b>29.91</b>	<b>28.72</b>	<b>28.30</b>

3, we report zero-shot clean and robust accuracy. We consider three types of adversaries ( $\epsilon_1 = 2/255$ ): 20-step PGD [33], CW [4], and AutoAttack (AA) [6]. Our MOO-AD effectively generalizes to diverse VLM architectures.

**Robustness transfer w.r.t. stronger adversaries.** We here extend our evaluations to larger perturbation radius  $\epsilon_1$  in Table 4. Specifically, we provide the average robust accuracy against adversaries of perturbation radii  $\epsilon_1 = 3/255$  &  $4/255$  across 15 datasets. Our MOO-AD stays robust even when faced with stronger unforeseen adversaries.

**Robustness transfer w/ PEFT.** Adversarial learning with full fine-tuning significantly improves zero-shot robustness but incurs high computational costs, particularly for large

Table 4. Comparison of MOO-AD and adversarial fine-tuning on average robustness on 15 datasets against larger perturbations  $\epsilon_1$ .

Radius $\epsilon_1$	Method	PGD	CW	AA
3/255	TeCoA [34]	17.90	17.51	17.08
	PMG-FT [43]	19.14	18.69	18.22
	FARE [38]	19.31	18.84	18.47
	MOO-AD	<b>25.16</b>	<b>25.02</b>	<b>24.29</b>
4/255	TeCoA [34]	11.23	10.70	10.35
	PMG-FT [43]	12.05	11.56	11.17
	FARE [38]	12.42	11.93	11.52
	MOO-AD	<b>18.47</b>	<b>17.11</b>	<b>16.64</b>

Table 5. Comparison of MOO-AD and adversarial fine-tuning on average performance ( $\epsilon_1 = 2/255$ ) on 15 datasets using VPT.

VPT-Extension	Clean	PGD	CW	AA
TeCoA [34]	42.61	18.12	16.88	15.39
PMG-FT [43]	42.11	19.26	17.68	16.47
FARE [38]	42.81	18.98	17.46	16.35
MOO-AD	<b>44.28</b>	<b>21.39</b>	<b>20.06</b>	<b>18.75</b>

VLMs. To mitigate this, we adopt Visual Prompt Tuning (VPT) [28], a PEFT strategy that optimizes a small set of trainable parameters within the token embedding layer. Table 5 reports average clean and robust accuracy across 15 datasets in the zero-shot setting. Even under parameter-

Table 6. Comparison of MOO-AD and adversarial fine-tuning on average robustness against **text-level and image-text adversaries**.

Method	Text-Level Attacks		Image-Text Attacks	
	BERT-Attack	GBDA	Co-Attack	VLAttack
TeCoA [34]	37.14	35.30	26.73	22.60
PMG-FT [43]	37.61	36.46	28.11	20.81
FARE [38]	35.45	34.97	25.38	23.15
<b>MOO-AD</b>	<b>40.72</b>	<b>40.18</b>	<b>30.24</b>	<b>28.07</b>

Table 7. BLIP-extension: Comparison of MOO-AD and adversarial fine-tuning in **image-text retrieval and image captioning**.

Method	Image-Text Retrieval				Image Captioning	
	Clean TR	Robust TR	Clean IR	Robust IR	Clean CIDEr	Robust CIDEr
TeCoA [34]	87.5	54.4	77.0	47.5	96.9	57.8
PMG-FT [43]	87.8	55.6	77.9	48.2	97.5	58.2
FARE [38]	88.2	55.9	78.4	49.0	98.1	58.7
<b>MOO-AD</b>	<b>90.8</b>	<b>58.2</b>	<b>80.6</b>	<b>51.3</b>	<b>99.3</b>	<b>61.5</b>

Table 8. Medical CLIP-extensions: Comparison of MOO-AD and adversarial fine-tuning in radiology imaging by the AUC score.

Method	ChestXray14		CheXpert		PadChest	
	Clean	PGD	Clean	PGD	Clean	PGD
TeCoA [34]	0.674	0.526	0.857	0.685	0.602	0.483
PMG-FT [43]	0.692	0.538	0.850	0.688	0.619	0.495
FARE [38]	0.687	0.533	0.845	0.679	0.615	0.490
<b>MOO-AD</b>	<b>0.724</b>	<b>0.561</b>	<b>0.883</b>	<b>0.715</b>	<b>0.632</b>	<b>0.527</b>

efficient optimization, MOO-AD outperforms prior adversarial fine-tuning methods by a large margin.

**Text-level and image-text adversarial robustness.** Beyond image-level robustness, we evaluate resilience to text-level and image-text adversaries. Specifically, we investigate **text-level adversaries** via BERT-Attack [31] and GBDA [24], alongside **image-text adversaries**, utilizing Co-Attack [50] and VLAttack [47] (Table 6). MOO-AD significantly improves zero-shot robustness against both attack types. We attribute this enhanced image-text robustness to adversarial refinement of image representations, which strengthens the robust image-text embedding space.

### 4.3. Image-Text Understanding & Medical Imaging

**Image-text understanding (BLIP extensions).** Unlike standard CLIP, BLIP employs a bootstrapped pre-training scheme to unify vision-language tasks. We evaluate **image-text retrieval** on Flickr30k [36] and **image captioning** on Nocaps [1]. Details are in Appendix A.2. Table 7 shows that MOO-AD outperforms others in natural performance and robustness. This highlights MOO-AD’s ability to inherit robust generalization from large VLMs, even in vision-language understanding.

**Medical imaging (Medical CLIP extensions).** We evaluate our MOO-AD in medical imaging, where adversaries pose critical risks to diagnosis. Following CheXzero [42],

Table 9. Ablation analysis of loss components and their variants in MOO-AD for average clean and robust accuracy on 15 datasets.

	MOPA		GDM		Clean	PGD	AA
	Untargeted	MOO	Untargeted	MOO			
1	✓				55.48	32.37	31.06
2	✓		✓		56.03	32.96	31.44
3		✓			57.72	34.43	33.29
		✓		✓	<b>58.96</b>	<b>35.70</b>	<b>34.16</b>

Table 10. Comparison of adversary types in MOO-AD, with OOD evaluation averaged over SUN397, Flower102, and CIFAR-100.

Adversary Type	ImageNet			Out-Of-Distribution		
	Clean	PGD	AA	Clean	PGD	AA
Untargeted Adversaries	53.42	29.13	28.05	51.14	16.87	16.20
Targeted Adversaries	50.94	25.88	25.34	53.27	20.59	19.63
<b>MOO-Adversaries</b>	<b>59.28</b>	<b>36.58</b>	<b>35.72</b>	<b>55.74</b>	<b>24.74</b>	<b>24.05</b>

we adopt a radiology-specific CLIP model and report the AUC scores on clean and adversarial (PGD-20,  $\epsilon_1 = 1/255$ ) medical data on ChestX-ray14 [44], CheXpert [27], and PadChest [3]. Our MOO-AD exhibits significant improvement in both natural performance and adversarial robustness across all medical benchmarks (Table 8), even on PadChest with rare clinical cases. Such a cross-domain extension further justifies the generalization capability of our MOO-AD method. More details are in Appendix A.2.

## 5. Analyses

### 5.1. Ablation Studies

**Effect of individual loss components.** Here, we investigate two main components of our MOO-AD: MOPA in Eq. (9) and GDM in Eq. (10) alongside their variants using untargeted adversaries. We report the average clean and robust accuracy w.r.t. diverse configurations in Table 9. Incorporating group-wise information (statistics) matching between teacher and student VLMs contributes to distributional robustness transfer, leading to improved zero-shot adversarial robustness without compromising natural performance. In addition, using MOO-adversaries instead of standard adversaries helps elicit comprehensive knowledge (better representations of the decision surface) from the teacher VLM, resulting in enhanced clean and robust accuracy.

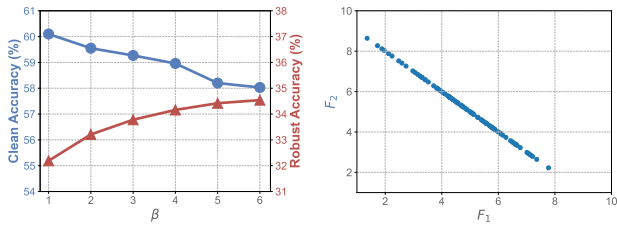
**Effect of diverse adversary types for distillation.** According to Figure 1, untargeted adversaries enhance in-distribution robustness, while targeted adversaries improve out-of-distribution (zero-shot) robustness. Table 10 presents the corresponding quantitative results on ImageNet (in-distribution) and the average performance across SUN397, Flower102, and CIFAR-100 (out-of-distribution). Our MOO-AD framework, leveraging MOO-adversaries, achieves superior robustness on both clean and adversarial examples in both settings, further validating our motivation.

Table 11. Comparison of different numbers of MOO-adversaries in MOO-AD for average clean and robust accuracy on 15 datasets with the average training time (minutes) per epoch.

MOO-Adversary Quantity	Clean	PGD	AA	Time (min)
$n = 1$ (Batch-Level Diversity)	56.97	32.17	30.89	54.8
$n = 5$	58.23	34.28	32.74	78.2
$n = 10$ (Our Setup)	58.96	35.70	34.16	98.9
$n = 20$	59.31	35.93	34.38	140.7
$n = 30$	59.47	36.11	34.49	181.5

Table 12. Comparison between boundary risk gain  $\kappa$  and natural risk  $\mathcal{R}_{nat}(\mathcal{D})$  w.r.t. MOO-AD on ImageNet across epochs.

Risk Metric	6-th	7-th	8-th	9-th	10-th
$\kappa$	0.391	0.362	0.325	0.295	0.286
$\mathcal{R}_{nat}(\mathcal{D})$	0.413	0.359	0.313	0.277	0.262
$\Delta$	-0.022	0.003	0.012	0.018	0.024



(a) Trade-off factor  $\beta$  (b) Objective Space Visualization.

Figure 3. (a) Accuracy-robustness trade-off by tuning  $\beta$ . (b) Adversary visualization in the objective space w.r.t.  $F_1$  and  $F_2$ .

### Impact of MOO-Adversary Quantity in Distillation

We examine whether increasing the number of MOO-adversaries improves zero-shot adversarial robustness by enhancing adversarial diversity. Specifically, we report the performance of MOO-AD under varying numbers of MOO-adversaries used during distillation (Table 11). For  $n = 1$ , adversaries are generated at the batch level, ensuring category diversity across the batch. For  $n > 1$ , diverse MOO-adversaries are generated per clean sample. While increasing adversary quantity improves robustness, the computational cost scales linearly, with performance converging at  $n = 10$ . Considering this performance-efficiency trade-off, we set  $n = 10$  during distillation. Note that our robust VLM shares the same inference time with other approaches.

**Comparison between  $\kappa$  and  $\mathcal{R}_{nat}(\mathcal{D})$ .** According Theorem 3, if the boundary risk gain  $\kappa$  is less than the natural risk  $\mathcal{R}_{nat}(\mathcal{D})$ , additionally introducing MOO-adversaries  $\mathcal{M}_x$  into  $\mathcal{D}$  may violate the minimization property of the robust risk’s upper bound. Table 12 compares  $\kappa$  against  $\mathcal{R}_{nat}(\mathcal{D})$  to determine whether the boundary risk gain offsets the natural risk. Results show that the assertion  $\kappa \geq \mathcal{R}_{nat}(\mathcal{D})$  holds throughout training, with the gap between the two metrics widening over adversarial knowledge distillation epochs.

### 5.2. Trade-offs (Hyper-parameter Analyses)

**Natural performance and adversarial robustness.** The trade-off between clean and robust accuracy is well-studied

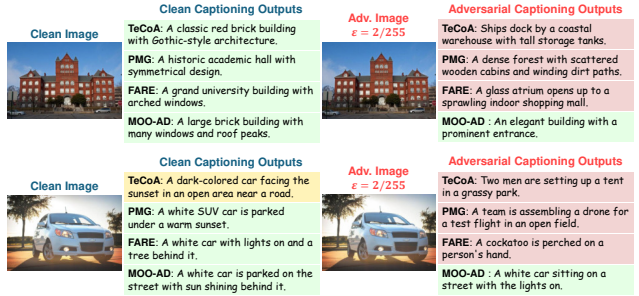


Figure 4. Robust BLIP Captioning on Nocaps of diverse methods.

in single-modal backbones but remains underexplored in VLMs. We investigate the balancing factor  $\beta$  in Eq.(9) that controls the weighting between clean and adversarial prediction alignment. Figure 3a reports clean and robust accuracy of different values of  $\beta$ . The robustness is enhanced when we enlarge  $\beta$  at the cost of natural performance.

**Category-level diversity.** We demonstrate that MOO-AD preserves diversity in the objective space. A randomly selected dataset batch is used to generate MOO-AD adversarial samples, visualized in Figure 3b w.r.t.  $F_1$  and  $F_2$ . As observed, the generated adversarial samples exhibit varying  $F_1$  values, and the entire batch is dispersedly distributed, effectively approximating the Pareto front of the multi-objective problem. As mentioned in Section 3.1,  $F_1$  represents the expected predicted class of a sample  $\hat{x}$ , indicating that diversity in the objective space aligns with category-level diversity. More explanations are in Appendix C.

### 5.3. Qualitative Visualizations

Beyond quantitative results, we present qualitative visualizations of adversarial examples in image captioning on Nocaps [1] using BLIP (Figure 4). MOO-AD generates high-quality, semantically accurate captions, even under adversarial perturbations, whereas other methods degrade significantly, often producing irrelevant or incoherent captions.

**Appendices.** A discusses experimental setups. B & C provide more background and explanations of MOO. Proofs of theorems are in E. Further analyses of our method are in F.

## 6. Conclusion

We identified the trade-off between in-distribution and out-of-distribution robustness in distillation, linking it to variations in adversarial diversity. To address this, we propose an adversarially robust knowledge distillation framework leveraging multi-objective optimization to generate adversaries with both diversity and disruptive capability, enabling a comprehensive exploration of the decision surface for robustness transfer. Theoretical analyses demonstrate the category-level diversity of multi-objective adversaries and show that their integration minimizes an upper bound on robust risk. Extensive experiments validate our method’s state-of-the-art robustness across diverse scenarios.

## Acknowledgments

This research is supported by National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative, 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.

## References

- [1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019. [7](#), [8](#)
- [2] J. Blank, K. Deb, Y. Dhebar, S. Bandaru, and H. Seada. Generating well-spaced points on a unit simplex for evolutionary many-objective optimization. *IEEE Transactions on Evolutionary Computation*, In press. [4](#)
- [3] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020. [7](#)
- [4] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017. [6](#)
- [5] Junxi Chen, Junhao Dong, and Xiaohua Xie. Mind the trojan horse: Image prompt adapter enabling scalable and deceptive jailbreaking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23785–23794, 2025. [2](#)
- [6] 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. [6](#)
- [7] Francesco Croce, Maksym Andriushchenko, Vikash Sehrawag, 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. [3](#)
- [8] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 958–979, 2024. [1](#)
- [9] 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. [5](#)
- [10] Natalia Díaz-Rodríguez, Javier Del Ser, Mark Coeckelbergh, Marcos López de Prado, Enrique Herrera-Viedma, and Francisco Herrera. Connecting the dots in trustworthy artificial intelligence: From ai principles, ethics, and key requirements to responsible ai systems and regulation. *Information Fusion*, 99:101896, 2023. [1](#)
- [11] 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, 2023. [2](#)
- [12] Junhao Dong, Yuan Wang, Jianhuang Lai, and Xiaohua Xie. Restricted black-box adversarial attack against deepfake face swapping. *IEEE Transactions on Information Forensics and Security*, 18:2596–2608, 2023. [2](#)
- [13] 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](#)
- [14] Junhao Dong, Junxi Chen, Xiaohua Xie, Jianhuang Lai, and Hao Chen. Survey on adversarial attack and defense for medical image analysis: Methods and challenges. *ACM Computing Surveys*, 57(3):1–38, 2024. [2](#)
- [15] 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. [1](#)
- [16] 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. [5](#)
- [17] Junhao Dong, Piotr Koniusz, Junxi Chen, Xiaohua Xie, and Yew-Soon Ong. Adversarially robust few-shot learning via parameter co-distillation of similarity and class concept learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28535–28544, 2024. [1](#)
- [18] 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. [1](#)
- [19] Junhao Dong, Melvin Wong, Sihan Xia, and Joel Wei En Tay. Towards adversarially robust data-efficient learning with generated data. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pages 1422–1424. IEEE, 2024. [2](#)
- [20] 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*, pages 236–247, 2025. [1](#)
- [21] Junhao Dong, Piotr Koniusz, Yifei Zhang, Hao Zhu, Weiming Liu, Xinghua Qu, and Yew-Soon Ong. Improving zero-shot adversarial robustness in vision-language models by

- closed-form alignment of adversarial path simplices. In *Forty-second International Conference on Machine Learning*, 2025. 1
- [22] Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3996–4003, 2020. 1, 2
- [23] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR*, 2015. 2
- [24] Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. Gradient-based adversarial attacks against text transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 5747–5757, 2021. 7
- [25] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2, 4
- [26] Bo Huang, Mingyang Chen, Yi Wang, Junda Lu, Minhao Cheng, and Wei Wang. Boosting accuracy and robustness of student models via adaptive adversarial distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24668–24677, 2023. 2
- [27] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597, 2019. 7
- [28] 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. 2, 6
- [29] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 1, 2
- [30] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27831–27840, 2024. 1
- [31] Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. BERT-ATTACK: adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing EMNLP*, pages 6193–6202, 2020. 7
- [32] 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. 1
- [33] 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. 3, 5, 6
- [34] 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. 1, 2, 3, 5, 6, 7
- [35] R Timothy Marler and Jasbir S Arora. The weighted sum method for multi-objective optimization: new insights. *Structural and multidisciplinary optimization*, 41:853–862, 2010. 2
- [36] 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. 7
- [37] 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, 5, 6
- [38] Christian Schlarman, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust CLIP: unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In *Forty-first International Conference on Machine Learning, ICML 2024*, 2024. 2, 5, 6, 7
- [39] Wu Song, Yong Wang, Han-Xiong Li, and Zixing Cai. Locating multiple optimal solutions of nonlinear equation systems based on multiobjective optimization. *IEEE Transactions on Evolutionary Computation*, 19(3):414–431, 2014. 4
- [40] David Stutz, Matthias Hein, and Bernt Schiele. Confidence-calibrated adversarial training: Generalizing to unseen attacks. In *International conference on machine learning*, pages 9155–9166. PMLR, 2020. 5
- [41] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014. 1
- [42] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406, 2022. 1, 7
- [43] Sibowang, 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. 1, 2, 5, 6, 7
- [44] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 7

- [45] Yong Wang, Han-Xiong Li, Gary G Yen, and Wu Song. Mommop: Multiobjective optimization for locating multiple optimal solutions of multimodal optimization problems. *IEEE transactions on cybernetics*, 45(4):830–843, 2014. 4
- [46] Mang Ye, Xuankun Rong, Wenke Huang, Bo Du, Nenghai Yu, and Dacheng Tao. A survey of safety on large vision-language models: Attacks, defenses and evaluations. *arXiv preprint arXiv:2502.14881*, 2025. 2
- [47] Ziyi Yin, Muchao Ye, Tianrong Zhang, Tianyu Du, Jinguo Zhu, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models. *Advances in Neural Information Processing Systems*, 36:52936–52956, 2023. 7
- [48] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. Dolphinattack: Inaudible voice commands. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pages 103–117, 2017. 2
- [49] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pages 7472–7482. PMLR, 2019. 5
- [50] Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5005–5013, 2022. 1, 2, 7
- [51] Qingfu Zhang and Hui Li. MOEA/D: A multiobjective evolutionary algorithm based on decomposition. *IEEE Transactions on evolutionary computation*, 11(6):712–731, 2007. 4
- [52] Xueyan Zhang, Jinman Zhao, Zhifei Yang, Yibo Zhong, Shuhao Guan, Linbo Cao, and Yining Wang. UORA: Uniform orthogonal reinitialization adaptation in parameter efficient fine-tuning of large models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 11709–11728. Association for Computational Linguistics, 2025. 2
- [53] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36:54111–54138, 2023. 2
- [54] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022. 2
- [55] Eckart Zitzler and Lothar Thiele. Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. *IEEE transactions on Evolutionary Computation*, 3(4):257–271, 1999. 2