Confound from All Sides, Distill with Resilience: Multi-Objective Adversarial Paths to Zero-Shot Robustness
– *Supplementary Material* –

Junhao Dong[1,2], Jiao Liu[1], Xinghua Qu[3], and Yew-Soon Ong[1,2‡]
[1]Nanyang Technological University, [2]CFAR, IHPC, A*STAR, [3]Bytedance
{junhao003, jiao.liu, asysong}@ntu.edu.sg, xinghua.qu1@bytedance.com

Abstract

In this supplementary material, we commence by providing a comprehensive description of our experimental setups (Appendix A), including dataset descriptions and implementation/extension details. Further, we introduce more background of Multi-Objective Optimization (MOO) in Appendix B for a better understanding. Additional explanations regarding our build MOO problem with intuitive examples are presented in Appendix C. The proofs of theoretical analyses are provided in Appendix E. Further Analyses of our MOO-AD method are in Appendix F.

A. Detailed Experimental Setups

A.1. Dataset Descriptions and Pre-Processing

In line with the evaluation protocols established in prior works [25, 33], we conduct adversarially robust knowledge distillation on the ImageNet training set [8] and assess its in-distribution robustness on the ImageNet validation set, commonly used as a test benchmark. For zero-shot (out-of-distribution) robustness evaluations, we test on additional 14 datasets that span diverse image recognition tasks: STL-10 [5], CIFAR-10/100 [16], Caltech-101/256 [10, 11], FGVC [24], Flower102 [28], Food101 [2], OxfordPets [29], and StandfordCars [15], DTD [4], EuroSAT [12], PCAM [36], and SUN397 [38]. For data pre-processing, input images are resized to 224×224 (except for CIFAR-10/100 and STL-10) and undergo center cropping before processing.

In addition to zero-shot classification, we further extend our MOO-AD method to vision-language understanding and medical image analysis. Specifically, we focus on the Flickr dataset [30] for bidirectional image-text retrieval and the Nocaps dataset [1] for image captioning. For medical image analysis, we conduct robustness evaluations on three standard multi-label radiology datasets: ChestX-ray14 [37], CheXpert [13], and PadChest [3].

*Corresponding author.
‡Corresponding author.

A.2. Further Implementation Detials

Standard configurations. For adversarially robust knowledge distillation, we adopt the CLIP architecture [31] with ViT-L/14 as the teacher model and ViT-B/32 & ResNet-50/101 as student models. The teacher VLM is obtained through the standard adversarial fine-tuning, *i.e.*, TeCoA [25]. Following [33], we fully optimize the vision encoder's parameters of the student VLM using AdamW [22] with the learning rate initialized at 1×10^{-5} using a cosine decay schedule for 10 epochs. In the case of Visual Prompt Tuning (VPT) [14], an efficient fine-tuning strategy, we incorporate 100 learnable tokens into the vision module of CLIP, setting the learning rate to 40. MOO-adversaries are generated with 10 iterations under the ℓ_∞ threat model with the radius $\epsilon_{\mathrm{I}} = 2/255$. The MOO weighting factors are set $\gamma_1 = \gamma_2 = 1.0$ for balanced optimization. Additionally, the loss weighting coefficients are $\lambda = 2.5$ and $\beta = 4.0$. All the hyper-parameter configurations are searched on a 10% subset of CIFAR-10 and then applied directly to adversarial distillation using ImageNet across diverse settings.

BLIP extension configurations. For evaluation metrics, we use recall@1 for Text Retrieval (TR) and Image Retrieval (IR) for both clean and adversarial examples. In the context of image captioning evaluations, CIDEr measures the similarity of a generated sentence against a set of ground truth sentences written by humans. We focus on adversarial examples of the perturbation radius $\epsilon_{\mathrm{I}} = 1/255$ generated by 20-step PGD attacks [23] during both training and evaluations, using the references/captions as labels. In line with [18], we directly integrate the Image-Text Contrastive (ITC) loss into our MOO-AD. For other adversarial fine-tuning approaches, we additionally incorporate the adversarial optimization of the ITC loss, the Image-Text Matching (ITM) loss, and the Language Modeling (LM) loss. We focus on the ViT-B/16-based BLIP architecture for evaluations.

Medical CLIP extension configurations. The Medical CLIP expansion follows CheXzero [35] by leveraging a radiology-specific CLIP model with a ViT/B-16 backbone. Note that the text encoder is replaced by BioBERT [17], a specialized biomedical language model optimized for text mining in medical scenarios. During the adversarial learning/distillation stage, we utilize a comprehensive chest X-ray benchmark including detailed radiology reports. At the inference stage, we evaluate the robust VLMs on ChestX-ray14 [37], CheXpert [13], and PadChest [3]. We report the Area Under the Curve (AUC) metric for both legitimate medical data and their adversarial counterparts (PGD-20, $\epsilon_{\mathrm{I}} = 1/255$).

Repulsive term in the MOO solver. To further keep the diversity of the MOO adversaries, we also add a repulsive potential term [32] into the generation process of the MOO-adversaries, which is

$$R(\hat{\mathbf{x}}_{\mathrm{MOO}}) = \sum_{m=1}^{N_b} \sum_{n=1}^{N_b} \exp\left(-\frac{||\mathbf{F}'(\hat{\mathbf{x}}_{\mathrm{MOO}}^{(m)}) - \mathbf{F}'(\hat{\mathbf{x}}_{\mathrm{MOO}}^{(n)})||}{\sigma^2}\right), \tag{14}$$

where $\mathbf{F}' = [F_1', F_2']$, and $\sigma = 0.05$ is a preset standard deviation. This repulsive term is combined with $\mathcal{L}^{\mathrm{tch}}\left(\hat{\mathbf{x}}_{\mathrm{MOO}}^{k,(i)}|\mathbf{w}^k\right)$ as part of the loss function to guide the iterative gradient ascent process.

B. Background of Multiobjective Optimization

Commonly, a MOO problem can be formulated as:

$$\begin{aligned} &\min : \mathbf{f}(\mathbf{x}) = \{f_1(\mathbf{x}), \dots, f_q(\mathbf{x})\}, \\ &\text{s.t. } \mathbf{x} \in \Omega \subseteq \mathbb{R}^d, \end{aligned} \tag{15}$$

where $f_l(\mathbf{x}), (l \in \{1, \dots, q\})$ is the lth objective function, q is the number of objectives, $\mathbf{x}$ is the decision vector, Ω is the decision space, d is the dimensions of the decision vector. Some key concepts associated with the MOO problem are introduced as follows [6]:

- *Pareto Dominance*: For decision vectors $\mathbf{x}_a$ and $\mathbf{x}_b$, if $\forall l \in \{1, 2, \dots, q\}$, $f_l(\mathbf{x}_a) \leq f_l(\mathbf{x}_b)$ and $\exists l' \in \{1, 2, \dots, q\}$, $f_{l'}(\mathbf{x}_a) < f_{l'}(\mathbf{x}_b)$, $\mathbf{x}_a$ is said to Pareto dominate $\mathbf{x}_b$.
- *Pareto Optimal Solution*: If no decision vector in Ω Pareto dominates $\mathbf{x}_a$, then $\mathbf{x}_a$ is a Pareto optimal solution.
- *Pareto Set*: The set of all Pareto optimal solutions forms the Pareto set in decision space.
- *Pareto Front*: The image of the Pareto set in the objective space forms the PF.

Unlike single-objective optimization, an MOO problem does not have a single solution that simultaneously minimizes or maximizes all objectives [6]. Instead, the goal is to identify a representative *set* of Pareto-optimal solutions that form the PF, representing the best achievable trade-offs in the objective space. Over the past few decades, MOO has been extensively studied in the optimization field. Evolutionary algorithms [7, 26, 27, 40–42] are a prominent class of methods for solving MOO problems. While their population-based search and inherent parallelism have proven effective, their inability to leverage gradient information limits their efficiency. Recently, gradient-based MOO optimizers have gained significant attention in various machine learning tasks [19, 21, 32, 34]. By incorporating gradient information, these methods enable efficient optimization in neural network-based problems, facilitating applications such as multi-task learning [34] and neural combinatorial optimization [20].

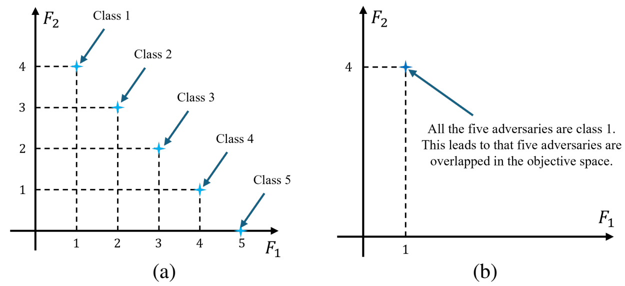

Figure 6. Examples of diversity preservation in the proposed multi-objective modeling approach. In the example, we show the representation of adversaries in the objective spaces corresponding to the constructed multi-objective optimization problem. We assuming there are two batches, $\mathcal{X}_1 = \{\hat{\mathbf{x}}_1^1, \dots, \hat{\mathbf{x}}_1^5\}$ and $\mathcal{X}_2 = \{\hat{\mathbf{x}}_2^1, \dots, \hat{\mathbf{x}}_2^5\}$, each consisting of five adversaries. In $\mathcal{X}_1$, the contained adversaries belong to five different classes, whereas in $\mathcal{X}_2$, all adversaries belong to a single class (class 1). Assume that for all adversaries in $\mathcal{X}_1$ and $\mathcal{X}_2$, their category membership is absolute, i.e., $[\mathbf{p}_{\mathcal{S}}(\hat{\mathbf{x}}')]_c = 1, c = \arg\max_c[\mathbf{p}_{\mathcal{S}}(\hat{\mathbf{x}}')]_c, \hat{\mathbf{x}}' \in \mathcal{X}_1 \cup \mathcal{X}_2$, and the constraints in (5) is fully guaranteed. (a) Representation of adversaries in the objective space for the batch $\mathcal{X}_1$. (b) Representation of adversaries in the objective space for the batch $\mathcal{X}_2$.

C. Additional Explanation on the Built MOO Problem

We present an example to intuitively illustrate diversity in the objective space in terms of adversarial samples. Consider two batches, $\mathcal{X}_1 = \{\hat{\mathbf{x}}_1^1, \dots, \hat{\mathbf{x}}_1^5\}$ and $\mathcal{X}_2 = \{\hat{\mathbf{x}}_2^1, \dots, \hat{\mathbf{x}}_2^5\}$, each containing five adversarial samples. In $\mathcal{X}_1$, the samples belong to five different classes, whereas in $\mathcal{X}_2$, all samples belong to a single class (class 1). Clearly, $\mathcal{X}_1$ exhibits greater class diversity than $\mathcal{X}_2$ in terms of class labels. Assuming the class membership of each sample is absolute ($[\mathbf{p}_{\mathcal{S}}(\hat{\mathbf{x}}')]_c = 1, c = \arg\max_c[\mathbf{p}_{\mathcal{S}}(\hat{\mathbf{x}}')]_c, \hat{\mathbf{x}}' \in \mathcal{X}_1 \cup \mathcal{X}_2$), and the constraints in Eq. (5) are fully satisfied, the expected predicted class of a sample $\hat{\mathbf{x}}$ is given by $F_1(\hat{\mathbf{x}}') = \arg\max_c[\mathbf{p}_{\mathcal{S}}(\hat{\mathbf{x}}')]_c$. Setting $C = 5$, the corresponding representations in the objective space are shown in Figure 6. As depicted in Figure 6a, $\mathcal{X}_1$, with high class-level diversity, also exhibits broad coverage of the Pareto front in the established MOO problem, ensuring diversity in the objective space. In contrast, all five adversaries in $\mathcal{X}_2$ overlap in the objective space, as shown in Figure 6b, demonstrating that insufficient class diversity leads to a failure in maintaining diversity in the objective space. This example highlights that in the formulated MOO problem, preserving diversity in the objective space corresponds to ensuring diversity among adversarial samples within a batch, considering class labels. Diversity preservation has been extensively studied in MOO, with well-established techniques. By integrating these strategies into our MOO formulation, the proposed method effectively maintains the diversity of adversarial samples.

In the above example, we primarily use samples from completely different classes to illustrate MOO-AD's ability to maintain diversity, and such examples may also be generated by targeted adversaries. However, since F_1 represents the expected likelihood that an adversarial sample belongs to a certain class, it is inherently a continuous value. By leveraging MOO, this continuous nature enables MOO-AD to generate adversarial samples that lie at the intersection of multiple decision boundaries (for example, $[\mathbf{p}_\mathcal{S}]_1 = 0.5$ and $[\mathbf{p}_\mathcal{S}]_2 = 0.5$). Consequently, adversarial samples generated by MOO-AD may provide more comprehensive coverage of the decision boundaries of student models compared to targeted adversaries, thereby contributing to the training of more robust models.

D. Robust Risk with MOO-Adversaries

D.1. Robust Risk Decomposition

Following [39], we decompose the robust risk $\mathcal{R}_{rob}$ into natural and boundary components. For a student VLM with predicted class $\mathbf{p}^*_\mathcal{S}(\mathbf{x}) = \text{argmax}_c[\mathbf{p}_\mathcal{S}(\mathbf{x})]_c$, the robust risk on a set $\mathcal{V}$ is defined as:

Definition 1 (Robust Risk [39])**.** *For sample-label pairs $(\mathbf{x}, c)$ drawn from $\mathcal{V}$, the robust risk and its two components—natural and boundary risks—are defined as follows:*

$$\begin{aligned}
&\mathcal{R}_{rob}(\mathbf{p}_\mathcal{S};\mathcal{V}) := \mathbb{E}_{(\mathbf{x},c)\sim\mathcal{V}}[\mathbb{1}(\exists\hat{\mathbf{x}} \in \mathbb{B}(\mathbf{x},\epsilon) : \mathbf{p}^*_\mathcal{S}(\hat{\mathbf{x}}) \neq c)],\\
&\mathcal{R}_{nat}(\mathbf{p}_\mathcal{S};\mathcal{V}) := \mathbb{E}_{(\mathbf{x},c)\sim\mathcal{V}}[\mathbb{1}(\mathbf{p}^*_\mathcal{S}(\mathbf{x}) \neq c)], \qquad (16)\\
&\mathcal{R}_{bdy}(\mathbf{p}_\mathcal{S};\mathcal{V}) := \mathbb{E}_{(\mathbf{x},c)\sim\mathcal{V}}[\mathbb{1}(\exists\hat{\mathbf{x}} \in \mathbb{B}(\mathbf{x},\epsilon) : \mathbf{p}^*_\mathcal{S}(\hat{\mathbf{x}}) \neq \mathbf{p}^*_\mathcal{S}(\mathbf{x}) = c)],
\end{aligned}$$

where ϵ is the ℓ_∞-norm perturbation radius around $\mathbf{x}$[1]*. Also, $\mathcal{R}_{rob}(\mathbf{p}_\mathcal{S};\mathcal{V}) = \mathcal{R}_{nat}(\mathbf{p}_\mathcal{S};\mathcal{V}) + \mathcal{R}_{bdy}(\mathbf{p}_\mathcal{S};\mathcal{V})$.*

D.2. Extension to MOO-Adversarial Examples

We then extend this decomposition to incorporate MOO-based adversarial examples generated via our proposed optimization strategy during robust knowledge distillation.

Definition 2. *Let $\mathcal{M}_\mathbf{x}$ denote a set of MOO-adversarial examples generated from each sample-label pair $(\mathbf{x}, c) \in \mathcal{D}$, i.e., $\mathcal{M}_\mathbf{x} = \cup_{(\mathbf{x},c)\in\mathcal{D}} \mathcal{M}_\mathbf{x}\big((\mathbf{x}, c)\big)$. We separate $\mathcal{M}_\mathbf{x}$ into two disjoint subsets: $\mathcal{M}^{✗}_\mathbf{x} = \{(\mathbf{x}, c) \in \mathcal{M}_\mathbf{x} : \mathbf{p}^*_\mathcal{S} \neq c\}$ and $\mathcal{M}^{✓}_\mathbf{x} = \{(\mathbf{x}, c) \in \mathcal{M}_\mathbf{x} : \mathbf{p}^*_\mathcal{S} = c\}$ that contain incorrectly and correctly classified MOO-adversaries, respectively.*

[1] Note that the set $\mathcal{V}$ may consist of both clean and adversarial data, hence the sample-label pair $(\mathbf{x}, c)$ can represent either type.

D.3. Robust Risk Bound Minimization

Building upon the robust risk bound analysis of intermediate adversarial examples introduced by [9], we extend their theoretical framework to the MOO-adversarial setting by replacing intermediate adversaries with MOO-adversarial examples and deriving new bounds tailored to this scenario:

Theorem 2. *Let $\mathcal{D}\cup\mathcal{M}_\mathbf{x}$ denote the original dataset combined with the MOO-adversaries. The robust risk gap compared to using only the original dataset $\mathcal{D}$ is given as:*

$$\mathcal{R}_{rob}(\mathcal{D}\cup\mathcal{M}_\mathbf{x}) - \mathcal{R}_{rob}(\mathcal{D}) = \qquad (17)$$
$$\frac{|\mathcal{M}^{✗}_\mathbf{x}|(\mathcal{R}_{rob}(\mathcal{M}^{✗}_\mathbf{x}) - \mathcal{R}_{rob}(\mathcal{D}))}{|\mathcal{D}| + |\mathcal{M}_\mathbf{x}|} + \frac{|\mathcal{M}^{✓}_\mathbf{x}|(\mathcal{R}_{rob}(\mathcal{M}^{✓}_\mathbf{x}) - \mathcal{R}_{rob}(\mathcal{D}))}{|\mathcal{D}| + |\mathcal{M}_\mathbf{x}|}$$

Proof. See Appendix E.2. □

Theorem 3. *Integrating MOO-adversaries into the robust risk $\mathcal{R}_{rob}(\mathcal{D}\cup\mathcal{M}_\mathbf{x})$ addresses an upper bound on the standard robust risk $\mathcal{R}_{rob}(\mathcal{D})$ of the original dataset $\mathcal{D}$,* i.e., *$\mathcal{R}_{rob}(\mathcal{D}\cup\mathcal{M}_\mathbf{x}) > \mathcal{R}_{rob}(\mathcal{D})$ given that $\kappa \geq \mathcal{R}_{nat}(\mathcal{D})$, where $\kappa = \mathcal{R}_{bdy}(\mathcal{M}^{✓}_\mathbf{x}) - \mathcal{R}_{bdy}(\mathcal{D}) \geq 0$ is the boundary-risk gap.*

Proof. See Appendix E.3. □

E. Theoretical Analyses

E.1. Proof of Theorem 1

Proof. According to $||\mathbf{F}(\mathbf{x}^a) - \mathbf{F}(\mathbf{x}^b)|| \geq \delta_\mathbf{F}$, we have

$$\begin{aligned}
&||\mathbf{F}(\mathbf{x}^a) - \mathbf{F}(\mathbf{x}^b)||\\
=&\sqrt{(F_1(\mathbf{x}^a) - F_1(\mathbf{x}^b))^2 + (F_2(\mathbf{x}^a) - F_2(\mathbf{x}^b))^2}\\
=&\sqrt{2(F_1(\mathbf{x}^a) - F_1(\mathbf{x}^b))^2} \geq \delta_\mathbf{F}.
\end{aligned}$$

Then, we can get $|F_1(\mathbf{x}^a) - F_1(\mathbf{x}^b)| \geq \frac{\sqrt{2}}{2}\delta_\mathbf{F}$. As F_1 is L-Lipschitz continuous, then we have $L||\mathbf{x}^a - \mathbf{x}^b|| \geq ||\mathbf{F}(\mathbf{x}^a) - \mathbf{F}(\mathbf{x}^b)|| \geq \frac{\sqrt{2}}{2}\delta_\mathbf{F}$. As a result, we get the corresponding conclusion $||\mathbf{x}^a - \mathbf{x}^b|| \geq \frac{\sqrt{2}\delta_\mathbf{F}}{2L}$. □

E.2. Proof of Theorem 2

Proof. By decomposing MOO-adversaries $\mathcal{M}_\mathbf{x}$ into $\mathcal{M}^{✗}_\mathbf{x} \cup \mathcal{M}^{✓}_\mathbf{x}$ based on the VLM classification, the robust risk gap can be expressed as an average of their respective robust risks, each weighted by its cardinality below:

$$\begin{aligned}
&\mathcal{R}_{rob}(\mathcal{D}\cup\mathcal{M}_\mathbf{x}) - \mathcal{R}_{rob}(\mathcal{D}) = \mathcal{R}_{rob}(\mathcal{D}\cup\mathcal{M}^{✗}_\mathbf{x}\cup\mathcal{M}^{✓}_\mathbf{x}) - \mathcal{R}_{rob}(\mathcal{D})\\
=&\frac{|\mathcal{D}|\,\mathcal{R}_{rob}(\mathcal{D}) + |\mathcal{M}^{✗}_\mathbf{x}|\,\mathcal{R}_{rob}(\mathcal{M}^{✗}_\mathbf{x}) + |\mathcal{M}^{✓}_\mathbf{x}|\,\mathcal{R}_{rob}(\mathcal{M}^{✓}_\mathbf{x})}{|\mathcal{D}| + |\mathcal{M}^{✗}_\mathbf{x}| + |\mathcal{M}^{✓}_\mathbf{x}|}\\
&- \frac{|\mathcal{D}| + |\mathcal{M}^{✗}_\mathbf{x}| + |\mathcal{M}^{✓}_\mathbf{x}|}{|\mathcal{D}| + |\mathcal{M}^{✗}_\mathbf{x}| + |\mathcal{M}^{✓}_\mathbf{x}|}\mathcal{R}_{rob}(\mathcal{D}) \qquad (18)\\
=&\frac{|\mathcal{M}^{✗}_\mathbf{x}|(\mathcal{R}_{rob}(\mathcal{M}^{✗}_\mathbf{x}) - \mathcal{R}_{rob}(\mathcal{D}))}{|\mathcal{D}| + |\mathcal{M}_\mathbf{x}|} + \frac{|\mathcal{M}^{✓}_\mathbf{x}|(\mathcal{R}_{rob}(\mathcal{M}^{✓}_\mathbf{x}) - \mathcal{R}_{rob}(\mathcal{D}))}{|\mathcal{D}| + |\mathcal{M}_\mathbf{x}|}.
\end{aligned}$$

Table 13. Comparison of diverse adversary mixing configurations in MOO-AD for average clean and robust accuracy on 15 datasets.

Adversary Mixing Configuration	Clean	PGD	AA
Untargeted & MOO Adversaries	56.48	33.25	32.08
Targeted & MOO Adversaries	57.65	34.18	32.83
Untargeted & Targeted & MOO Adversaries	58.15	34.79	33.42
MOO-Adversaries Only (**Ours**)	**58.96**	**35.70**	**34.16**

□

E.3. Proof of Theorem 3

Proof. To establish that $\mathcal{R}_{rob}(\mathcal{D}\cup\mathcal{M}_{\mathbf{x}})\geq\mathcal{R}_{rob}(\mathcal{D})$, it suffices to show two inequalities hold below: (i) $\mathcal{R}_{rob}(\mathcal{M}_{\mathbf{x}}^{\text{✗}})-\mathcal{R}_{rob}(\mathcal{D})\geq 0$ & (ii) $\mathcal{R}_{rob}(\mathcal{M}_{\mathbf{x}}^{\text{✓}})-\mathcal{R}_{rob}(\mathcal{D})\geq 0$.

We first address the condition (i). By Definition 1 of the robust and natural risks, any instance in $\mathcal{M}_{\mathbf{x}}^{\text{✗}}$ is misclassified by formulation; consequently, $\mathcal{R}_{rob}(g_\theta;\mathcal{M}_{\mathbf{x}}^{\text{✗}}) = \mathcal{R}_{nat}(g_\theta;\mathcal{M}_{\mathbf{x}}^{\text{✗}})=1\geq\mathcal{R}_{rob}(g_\theta;\mathcal{D}) \geq \mathcal{R}_{nat}(g_\theta;\mathcal{D})$. Following Definition 2, $\mathcal{R}_{nat}(g_\theta;\mathcal{M}_{\mathbf{x}}^{\text{✗}})=1$ as all clean samples of $\mathcal{M}_{\mathbf{x}}^{\text{✗}}$ are misclassified. Since boundary risk $\mathcal{R}_{bdy}$ requires correctly classified clean samples, it vanishes for the misclassified set. This remains consistent with $\mathcal{M}_{\mathbf{x}}^{\text{✗}}$ being entirely misclassified and thus confirms condition (i).

We next examine the condition (ii), *i.e.*, $\mathcal{R}_{rob}(\mathcal{M}_{\mathbf{x}}^{\text{✓}}) - \mathcal{R}_{rob}(\mathcal{D})$. By Definition 2, all the elements in $\mathcal{M}_{\mathbf{x}}^{\text{✓}}$ are correctly classified, thus the natural risk $\mathcal{R}_{nat}(\mathcal{M}_{\mathbf{x}}^{\text{✓}}) = 0$. However, the boundary risk $\mathcal{R}_{bdy}(\mathcal{M}_{\mathbf{x}}^{\text{✓}})$ can be nonzero, as small perturbations to correctly classified legitimate samples can shift them across the decision boundary. Let $\kappa=\mathcal{R}_{bdy}(\mathcal{M}_{\mathbf{x}}^{\text{✓}})-\mathcal{R}_{bdy}(\mathcal{D})\geq 0$ represent the boundary risk gain. Consequently, if $\kappa\geq\mathcal{R}_{nat}(\mathcal{D})$, we obtain the condition (ii). Putting parts (i) and (ii) together completes the argument, showing that $\mathcal{R}_{rob}(\mathcal{D}\cup\mathcal{M}_{\mathbf{x}})$ is an upper bound of the adversarially robust risk $\mathcal{R}_{rob}(\mathcal{D})$. □

F. Further Analyses of Our MOO-AD Method

F.1. Impact of Adversary Mixing in Distillation.

Beyond using the MOO-adversaries alone, we examine whether incorporating targeted and/or untargeted adversaries enhances robustness transfer. Table 13 presents the distillation results for different adversary mixing configurations in MOO-AD. Interestingly, we find that introducing additional (targeted/untargeted) adversaries during distillation deteriorates zero-shot adversarial robustness. We attribute this robustness degradation to the potential disruption of adversarial diversity, which implicitly compromises the effectiveness of robustness transfer.

F.2. Analyses of In-Distribution and Out-Of-Distribution Robustness.

We here analyze the inherent relationship between the preset weights (γ_1 and γ_2) in the MOO solver and the trade-off

Table 14. Comparison of $\gamma_1\&\gamma_2$ values in MOO-AD, with OOD evaluation averaged over SUN397, Flower102, and CIFAR-100.

$\gamma_1\&\gamma_2$ Values	ImageNet			Out-Of-Distribution		
	Clean	PGD	AA	Clean	PGD	AA
$\gamma_1=\gamma_2=0.5$	58.14	35.93	35.19	56.30	25.12	24.37
$\gamma_1=\gamma_2=1.0$	59.28	36.58	35.72	55.74	24.74	24.05
$\gamma_1=\gamma_2=2.0$	59.67	37.02	36.13	54.92	24.28	23.69

between the in-distribution and out-of-distribution robustness. To ensure a consistent data view across the teacher and student VLMs during distillation, we set $\gamma_1=\gamma_2$. According to Table 14, we report the performance in both in-distribution (ImageNet) and out-of-distribution (average over SUN397, Flower102, and CIFAR-100) scenarios. Typically, Increasing both γ_1 and γ_2 enhances the disruptive capability of adversaries, leading to better in-distribution robustness. On the other hand, reducing them facilitates a more diverse MOO-adversary generation, resulting in improved out-of-distribution adversarial robustness.

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. 12

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 446–461. Springer, 2014. 12

[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. 12, 13

[4] 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. 12

[5] 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. 12

[6] Kalyanmoy Deb. *Multi-objective optimisation using evolutionary algorithms: an introduction*. Springer, 2011. 13

[7] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Trans. on Evol. Comput.*, 6(2):182–197, 2002. 13

[8] 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. 12

[9] Junhao Dong, Piotr Koniusz, Junxi Chen, Z Jane Wang, and Yew-Soon Ong. Robust distillation via untargeted and tar-

geted intermediate adversarial samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28432–28442, 2024. 14

[10] 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. 12

[11] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007. 12

[12] 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. 12

[13] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana 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. 12, 13

[14] 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. 12

[15] 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. 12

[16] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 12

[17] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. 13

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 12

[19] Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qing-Fu Zhang, and Sam Kwong. Pareto multi-task learning. *Advances in neural information processing systems*, 32, 2019. 13

[20] Xi Lin, Zhiyuan Yang, and Qingfu Zhang. Pareto set learning for neural multi-objective combinatorial optimization. *arXiv preprint arXiv:2203.15386*, 2022. 13

[21] Suyun Liu and Luis Nunes Vicente. The stochastic multi-gradient algorithm for multi-objective optimization and its application to supervised machine learning. *Annals of Operations Research*, 339(3):1119–1148, 2024. 13

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR*, 2019. 12

[23] 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. 12

[24] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 12

[25] 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. 12

[26] Wali Khan Mashwani and Abdellah Salhi. A decomposition-based hybrid multiobjective evolutionary algorithm with dynamic resource allocation. *Applied Soft Computing*, 12(9): 2765–2780, 2012. 13

[27] Wali Khan Mashwani, Abdellah Salhi, Muhammad Sulaiman, Rashida Adeeb Khanum, Abdulmohsen Algarni, et al. Evolutionary algorithms based on decomposition and indicator functions: State-of-the-art survey. *International Journal of Advanced Computer Science and Applications*, 7 (2), 2016. 13

[28] 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. 12

[29] 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. 12

[30] 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. 12

[31] 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. 12

[32] Yinuo Ren, Tesi Xiao, Tanmay Gangwani, Anshuka Rangi, Holakou Rahmanian, Lexing Ying, and Subhajit Sanyal. Multi-objective optimization via wasserstein-fisher-rao gradient flow. In *International Conference on Artificial Intelligence and Statistics*, pages 3862–3870. PMLR, 2024. 13

[33] 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 *Forty-first International Conference on Machine Learning, ICML 2024*, 2024. 12

[34] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 525–536, Red Hook, NY, USA, 2018. Curran Associates Inc. 13

[35] 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. 13

[36] 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. 12

[37] 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. 12, 13

[38] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 12

[39] 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. 14

[40] Qingfu Zhang and Hui Li. Moea/d: A multiobjective evolutionary algorithm based on decomposition. *IEEE Trans. on Evol. Comput.*, 11(6):712–731, 2007. 13

[41] Eckart Zitzler and Simon Künzli. Indicator-based selection in multiobjective search. In *Parallel Problem Solving from Nat. - PPSN VIII*, pages 832–842, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[42] Eckart Zitzler, Marco Laumanns, and Lothar Thiele. Spea2: Improving the strength pareto evolutionary algorithm. *TIK-report*, 103, 2001. 13