

MMT-ARD: Multimodal Multi-Teacher Adversarial Distillation for Robust Vision-Language Models

Yuqi Li^{1*} Junhao Dong^{2*} Chuanguang Yang² Shiping Wen⁴
 Piotr Koniusz⁵ Tingwen Huang⁶ Yingli Tian^{1†} Yew-Soon Ong^{2†}

¹The City University of New York, CUNY

²Nanyang Technological University

³Institute of Computing Technology, Chinese Academy of Sciences

⁴University of Technology Sydney

⁵Data61, CSIRO

⁶Shenzhen University of Advanced Technology

Abstract

Vision-Language Models (VLMs) are increasingly deployed in safety-critical applications, making their adversarial robustness a crucial concern. While adversarial knowledge distillation has shown promise in transferring robustness from teacher to student models, traditional single-teacher approaches suffer from limited knowledge diversity, slow convergence, and difficulty in balancing robustness and accuracy. To address these challenges, we propose MMT-ARD: a Multimodal Multi-Teacher Adversarial Robust Distillation framework. Our key innovation is a dual-teacher knowledge fusion architecture that collaboratively optimizes clean feature preservation and robust feature enhancement. To better handle challenging adversarial examples, we introduce a dynamic weight allocation strategy based on teacher confidence, enabling adaptive focus on harder samples. Moreover, to mitigate bias among teachers, we design an adaptive sigmoid-based weighting function that balances the strength of knowledge transfer across modalities. Extensive experiments on ImageNet and zero-shot benchmarks demonstrate that MMT-ARD improves robust accuracy by +4.32% and zero-shot accuracy by +3.5% on the ViT-B-32 model, while achieving a 2.3× increase in training efficiency over traditional single-teacher methods. These results highlight the effectiveness and scalability of MMT-ARD in enhancing the adversarial robustness of multimodal large models. Our codes are available at <https://github.com/itsnotacie/MMT-ARD>

1. Introduction

With the rapid advancement of multimodal artificial intelligence technology, Vision-Language Models (VLMs) have been widely adopted in autonomous driving, medical imaging, and industrial inspection. By jointly learning visual and textual representations, these models demonstrate strong cross-modal reasoning abilities. However, VLMs remain highly vulnerable to adversarial perturbations. Studies show that adding imperceptible perturbations can lead to completely erroneous model predictions [13] such as traffic-sign misclassification in autonomous driving or diagnostic errors in medical settings. This fragility stems from the multimodal alignment mechanism of VLMs—attackers disrupt cross-modal attention calculations by perturbing critical regions in the visual feature space, causing the model to produce high-confidence erroneous matches for adversarial examples. As VLMs enter safety-critical applications, their adversarial vulnerability has emerged as a major security threat hindering technological deployment. To break through the Current mainstream defenses fall into three categories: adversarial training, parameter-efficient fine-tuning, and knowledge distillation. Adversarial training enhances robustness by minimizing adversarial loss, but is computationally expensive[1, 8, 19]. Parameter-efficient fine-tuning methods (e.g., prompt tuning) reduce computational requirements yet rely heavily on the inherent robustness of pre-trained models, leading to poor cross-dataset generalization[3, 14]. Knowledge distillation, particularly Adversarial Robustness Distillation (ARD), have shown great potential in enhancing model resilience. However, existing approaches still suffer from three key limitations: 1) Foundational fine-tuning flaw: they rely on fine-tuning non-robust large models as teachers, which is

*Equal Contribution.

†Corresponding Author.

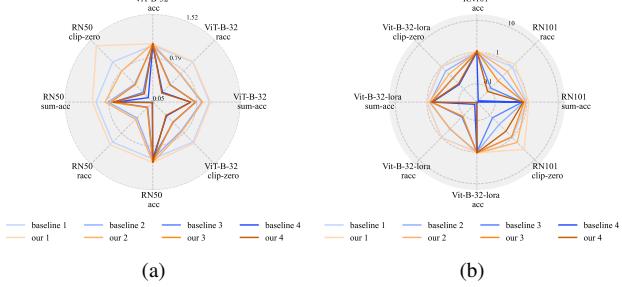


Figure 1. Multidimensional performance comparison of MMT-ARD with the baseline under different backbone. (a) Teacher-student combination based on ViT-B-32 and RN50. (b) Combination based on ViT-B-32-lora and RN101. The method proposed in this study (Our 1-4) comprehensively outperforms the baseline methods (Baseline 1-4) across the clean accuracy (acc) and robust accuracy (racc).

costly and ineffective in addressing the inherent structural vulnerabilities; 2) Convergence efficiency bottleneck: Student models require hundreds of epochs to approach teacher performance, making it difficult to meet practical deployment efficiency requirements; and 3) Single-teacher architecture limitation: A single teacher cannot simultaneously transfer both strong discriminative (clean) and robust (adversarial) features, resulting in an inevitable trade-off between clean accuracy and robustness. To overcome these limitations, we propose a Multimodal Multi-Teacher Adversarial Robustness Distillation (MMT-ARD) framework. The main contributions are summarized as follows: 1. A **multimodal multi-teacher knowledge fusion architecture** is designed to achieve synergistic optimization between clean feature preservation and robust feature enhancement. 2. A **Dynamic Importance Weighting (DIW) algorithm** is proposed to adaptively balance the knowledge transfer intensity from multiple teachers based on confidence and feature relevance. 3. A **cross-modal consistency constraint loss** is constructed to enhance adversarial invariance within the visual-textual embedding space, improving the model’s robustness under multimodal perturbations.

Extensive experiments on ImageNet and Zero-Shot benchmarks demonstrate the effectiveness of the proposed method, showing significantly improvements on ViT-B-32 robustness by 4.32%, zero-shot accuracy by 3.5%, and training efficiency by 2.3x over traditional adversarial distillation approaches. As shown in Figure 1. It can be clearly observed that under different architectures (such as ResNet and Vision Transformer), the performance polygon of our method ('Our') significantly encloses that of the baseline ('Baseline'), indicating that our method achieves overall performance improvements in clean accuracy, robust accuracy, and generalization metrics.

2. Related Work

2.1. Adversarial Attack

Adversarial attacks aim to mislead deep learning model by adding carefully crafted perturbations. Depending on the attacker’s level of knowledge about the target model, adversarial attack research has evolved into three categories: optimization-driven attacks (e.g., FGSM [17], PGD [11]) which iteratively optimize perturbations to maximize prediction errors; attention-reconstruction attacks (e.g., AOA [4], TAIG [9]) which manipulate the model’s attention maps to disrupt feature localization; and decision-smoothing attacks (e.g., TI [6], DI [20]) which improve transferability by smoothing the loss. Hybrid methods such as SM2I-FGSM [15] combine these strategies to exceed the limits of single-mechanism attacks. With the popularity of multimodal foundation models, adversarial research has expanded toward attacking multi-model cooperative systems.

2.2. Adversarial Robustness via Finetuning

Traditional single-modal defenses such as SAT [2] and TRADES [24] improve robustness through min-max optimization but fail under cross-modal attacks [10] and suffer significant drops in zero-shot generalization performance [12], which limits their utility in open environments. In contrast, multimodal cooperative defense offers a more systematic and resilient solution. Text-guided contrastive defenses (e.g., PMG-AFT) improve robustness by freezing the text encoder to stabilize the shared feature space [16], thus achieving robust accuracy gains on ImageNet. Meanwhile, cross-modal feature alignment methods (e.g., FARE) employ unsupervised adversarial fine-tuning, which eventually reduces the adversarial feature bias to below 0.1. More importantly, multimodal defense establishes a "cross-modal immune system" [13], which greatly improves the defense rate of joint attacks in scenarios such as payment systems. Collectively, these advances demonstrate that vision-language joint optimization effectively overcomes the cross-modal vulnerability of single-modal defense and provides a robust and generalizable protection mechanism for open environments.

2.3. Knowledge Distillation

The core framework of knowledge distillation [21–23] is to transfer valuable knowledge from the teacher to the student. Traditional Robust knowledge Distillation methods (such as RSLAD [28]) introduce robust soft labels but remains constrained by the single-teacher ceiling: student performance cannot surpass that of its teacher. The defense success rate under black-box attacks is still less than 50%. Traditional single-teacher adversarial robust distillation exposes the modal fragmentation predicament: visual teachers cannot guide text adversarial defense, result-

ing in fatal vulnerabilities in multimodal system defense. The multi-teacher knowledge distillation framework introduced to the study of adversarial distillation [25, 27]. It is worth noting that our research extends multi-teacher distillation to both robust and multimodal large language model contexts. The key intuition is that different robust teacher models (trained via distinct adversarial strategies) possess complementary strengths in handling various input regions or semantic attributes[7, 18, 26]. By allowing the student to learn collaboratively from multiple robust teachers, the proposed framework enables the integration of diverse robustness cues, producing student models that not only inherit but often surpass the robustness of any individual teacher.

3. Method

3.1. Multimodel Multi-Teacher Adversarial Robust Distillation

Inspired by multi-teacher and robust unsupervised finetuning, we propose the Multimodel Multi-Teacher Adversarial Robust Distillation (MMT-ARD) framework. The core idea of this method is to simultaneously utilize an Adversarial Teacher and a Clean Teacher to guide the training of a student CLIP model, thereby significantly improving the robustness of the model under adversarial attacks while maintaining the consistency of its multimodal embeddings. This design ensures consistent cross-modal feature representations while maintaining strong performance under both clean and adversarial conditions. The overall architecture of our proposed MMT-ARD framework is illustrated in Figure 2.

We employ the fine-tuned CLIP model as the adversarial teacher and the original CLIP model as the clean teacher. During training, the student model is jointly supervised by both teachers: the adversarial teacher provides a robust feature representation under adversarial samples, whose input is the adversarial samples generated when the student model is internally maximized, while the clean teacher provides a semantic feature representation under clean samples. The student model receives both adversarial and clean inputs, producing outputs that are guided by corresponding adversarial soft labels and clean soft labels. Therefore, the robustness optimization framework of the proposed MMT-ARD method can be formulated as follows:

$$O_{FT} = \arg \min \sum_{i=1}^n \left[(1 - \alpha) \cdot KL(S_{org}(x_i), T_{org}(x_i)) + \alpha \cdot KL(P_S(x_i), P_T(x_i)) \right], \quad (1)$$

$$P_m(x) = \max_{\delta \leq \varepsilon} \|m_{adv}(x + \delta) - m_{org}(x)\|_2^2, \quad (2)$$

where δ defines the perturbation constraint for generating adversarial samples, ensuring that the resulting perturbations are imperceptible to the human eye. Specifically, this constraint limits the pixel-wise change in an image to not exceed a small positive threshold ϵ , thereby preserving the visual appearance of the original input. Among them, S_{org} represents the clean student model, T_{org} represents the clean teacher model, m_{adv} represents the adversarial student model or the adversarial teacher model, and max represents the element with the largest absolute value within the feature space. The hyperparameter α controls the relative importance of the two sub-objectives in the final optimization process. By adjusting α , the training process can flexibly balance the emphasis between clean and adversarial objectives. In the following section, we introduce an adaptive parameter mechanism designed to dynamically regulate this weighting within the loss function.

3.2. Dynamic Weight of Teachers' Confidence

In the multi-teacher distillation framework, traditional static weight distribution methods exhibit two key limitations. First, the reliability of knowledge sources differs inherently between teachers. The adversarial teacher's prediction confidence for adversarial samples typically shows a bimodal distribution—where high-confidence correctly defended samples coexist with low-confidence attacked samples—whereas the clean teacher's confidence distribution on original samples is unimodal and more stable. Second, the weight distribution should have sample dependence: predictions for simple categories (e.g., “dog”) tend to be more confident than those for complex scenes (e.g., “crowded marketplace”). Static weighting, therefore, cannot adapt to the semantic complexity and difficulty of different samples. To address these issues, we propose a dynamic weight allocation strategy grounded in three core principles: 1) Deterministic priority principle: Assign higher weights to high-confidence predictions to ensure reliable knowledge transfer. 2) Uncertainty penalty principle: Suppress the interference of noise signals by reducing the weight of low-confidence predictions. 3) Cross-modal alignment principle: Promote the consistency of multimodal representations through the joint estimation of confidence across visual and linguistic modalities. This dynamic weight distribution strategy essentially builds a sample-adaptive knowledge fusion mechanism, enabling the model to automatically adjust the degree of trust assigned to different teachers based on specific sample features, thereby achieving more precise and robust knowledge distillation.

Definition of Teacher Confidence: Given a teacher model T and an input x , its prediction confidence is:

$$\text{conf}_T(x) = \max(\sigma(T(x))), \quad (3)$$

where $\sigma()$ denotes the softmax function and $T(x) \in \mathbb{R}^C$ is

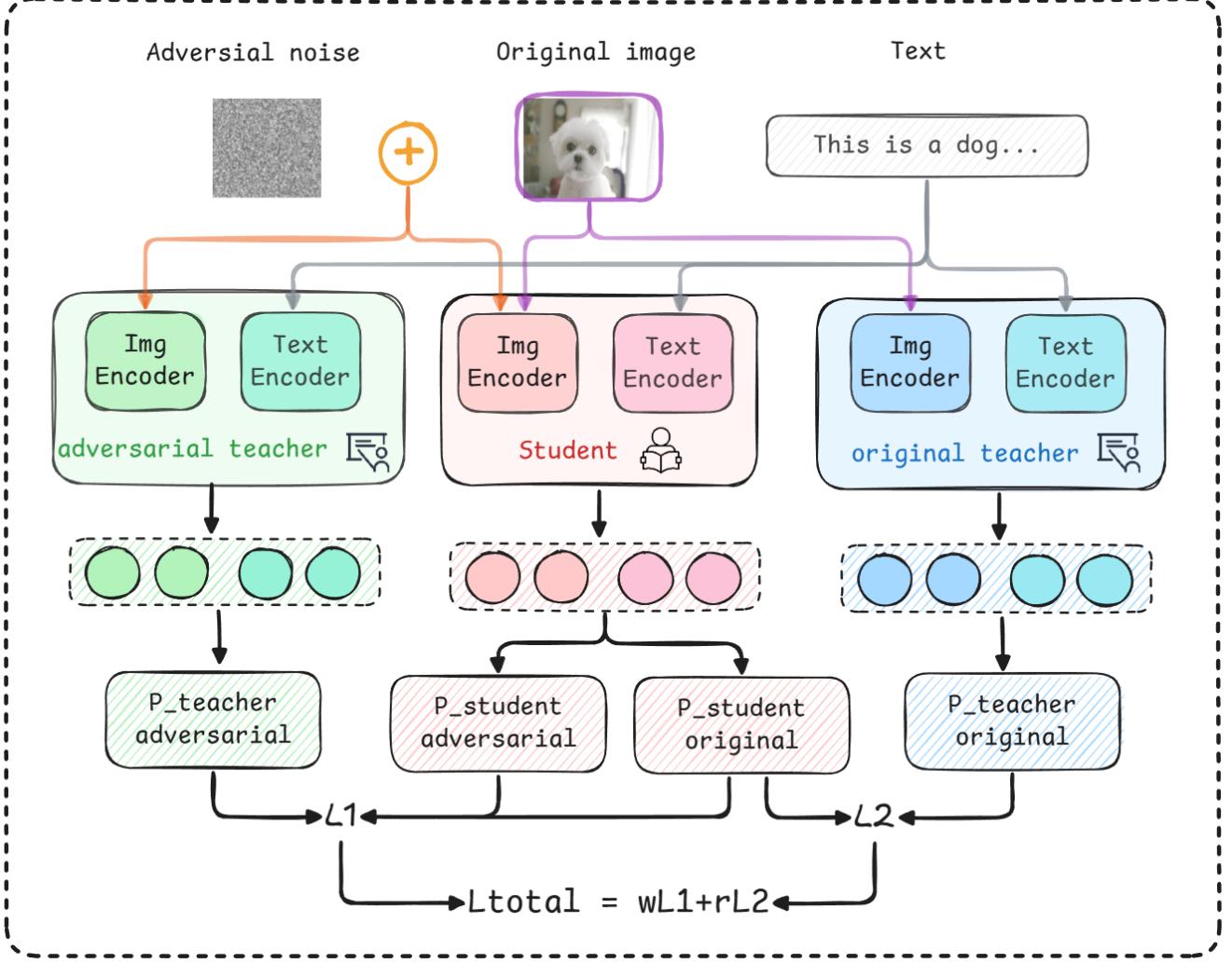


Figure 2. MMT-ARD framework architecture, where the same input image is processed separately by two sets of encoders from the original teacher and the adversarial teacher. L1 and L2, which respectively constrain the consistency of the student model’s outputs with those of the two teachers, ultimately achieving collaborative transfer of robust representations through a weighted sum.

the categorical logits vector. Dynamic weight calculation: for adversarial teacher T_{adv} and clean teacher T_{org} , define the weight ratio as follows.

$$\rho(x) = \frac{\text{conf}_{T_{adv}}(x)}{\text{conf}_{T_{org}}(x) + v}, \quad (4)$$

where $v = 10^{-5}$ is the numerically stable term. The final weights are generated using the modified sigmoid function.

$$w_{adv}(x) = \frac{1}{1 + e^{-\lambda(\rho(x)-\tau)}}, \quad (5)$$

$$w_{clean}(x) = 1 - w_{adv}(x), \quad (6)$$

where λ denotes the slope coefficient, which controls the sharpness of the weight change and τ is the offset to adjust

the weight balance.

4. Theoretical Analyses

Robustness Transfer under Multi-Teacher Distillation. Let $\{z^{(m)} : \mathcal{X} \rightarrow \mathbb{R}^K\}_{m=1}^M$ be M teacher logit maps with nonnegative weights w_1, \dots, w_M such that $\sum_m w_m = 1$. For a labeled input (x, y) , define the *teacher margins* $\gamma^{(m)}(x) := z_y^{(m)}(x) - \max_{k \neq y} z_k^{(m)}(x)$, $m = 1, \dots, M$, and the *logit-averaged ensemble* $z^{ens}(x) := \sum_{m=1}^M w_m z^{(m)}(x)$, $\Gamma_{ens}(x) := z_y^{ens}(x) - \max_{k \neq y} z_k^{ens}(x)$. Suppose that the student logit map $z^S : \mathcal{X} \rightarrow \mathbb{R}^K$ is L_S -Lipschitz w.r.t. ℓ_2 and fits the ensemble at x within ℓ_∞ discrepancy $\Delta(x) := \|z^S(x) - z^{ens}(x)\|_\infty$. Then for any perturbation δ with $\|\delta\|_2 \leq \varepsilon$, the

student's perturbed margin satisfies the following:

$$\underbrace{z_y^S(x + \delta) - \max_{k \neq y} z_k^S(x + \delta)}_{\text{student margin at } x + \delta} \geq \underbrace{\sum_{m=1}^M w_m \gamma^{(m)}(x)}_{\text{avg. teacher margin at } x} - 2\Delta(x) - 2L_S \varepsilon. \quad (7)$$

In particular, the student's top-1 prediction at $x + \delta$ remains y for all $\|\delta\|_2 \leq \varepsilon$ whenever

$$\varepsilon < \frac{\sum_{m=1}^M w_m \gamma^{(m)}(x) - 2\Delta(x)}{2L_S}.$$

Ensemble margin vs. average teacher margins. By convexity of \max , for any vectors $a^{(m)} \in \mathbb{R}^K$, $\max_k (\sum_m w_m a_k^{(m)}) \leq \sum_m w_m \max_k a_k^{(m)}$. With $a^{(m)} = z^{(m)}(x)$, we get $\Gamma_{\text{ens}}(x) \geq \sum_m w_m \gamma^{(m)}(x)$.

Student-ensemble closeness. From $\|z^S(x) - z^{\text{ens}}(x)\|_\infty \leq \Delta(x)$, $z_y^S(x) \geq z_y^{\text{ens}}(x) - \Delta(x)$ and $\max_{k \neq y} z_k^S(x) \leq \max_{k \neq y} z_k^{\text{ens}}(x) + \Delta(x)$, we have

$$z_y^S(x) - \max_{k \neq y} z_k^S(x) \stackrel{(1)}{\geq} \Gamma_{\text{ens}}(x) - 2\Delta(x) \geq \sum_m w_m \gamma^{(m)}(x) - 2\Delta(x).$$

Lipschitz stability. Since each logit of z^S is L_S -Lipschitz, $|z_c^S(x + \delta) - z_c^S(x)| \leq L_S \|\delta\|_2$ for all classes c . Thus the margin can shrink by at most $2L_S \|\delta\|_2$:

$$z_y^S(x + \delta) - \max_{k \neq y} z_k^S(x + \delta) \geq (z_y^S(x) - \max_{k \neq y} z_k^S(x)) - 2L_S \|\delta\|_2.$$

Combining Eqs. (2) and (3) gives the claim; the robustness condition follows by positivity of the right-hand side.

Remark 1. The theorem states the student inherits robustness from multiple teachers through their average margin, but loses some due to imperfect matching of the ensemble and sensitivity to input changes. To strengthen guarantees, increase teachers' margins, reduce the student-ensemble mismatch during distillation, and control the student's Lipschitz constant.

5. Experiment

5.1. Dataset Description

ImageNet-1K[5] serves as the main primary dataset for both training and evaluation, where adversarial samples are generated to assess model robustness. Additionally, we evaluate the model's generalization capability on zero-shot classification tasks following the standard zero-shot evaluation protocol of the CLIP pre-trained model.

5.2. Implementation Details

Model architecture and training configuration: For model selection, we used OpenFlamingo 9 B and LLaVA-1.5 7 B as LVLM models as the infrastructure for teacher

and student models. For the Teacher model, we choose dual teachers (adversarial teacher and Clean teacher), including ViT-L-14_PMG_Fast2 (adversarial training version and self-trained Clean Teacher (based on ViT-L-14). On the student model, we experiment with four networks, including ViT-B-32, RN50, RN101, ViT-B-32-LoRA (using the LoRA fine-tuning strategy). All experiments were performed under the same hardware environment (NVIDIA A100), and the results were repeated three times and averaged to ensure statistical significance.

Evaluation metrics include: Clean Accuracy (acc) : The classification accuracy of the model on clean samples. Robust Accuracy (racc): The classification accuracy of the model on adversarial samples. Adversarial samples are generated using PGD attacks, with attack intensities (ϵ) of 1/255, 2/255, 3/255, and 4/255, respectively. Sum-ACC: The Sum of clean accuracy and robust accuracy, which is used to comprehensively evaluate the model performance. Zero-Shot Accuracy: Accuracy on zero-shot classification tasks.

5.3. Comprehensive Comparative Experiments on MM-TARD

5.3.1. Enhanced robustness

As shown in Table 1, under low-intensity attack scenarios, our method achieves a 4.32% absolute improvement in robust accuracy (racc) over the baseline (from 45.02% to 49.34%), representing a statistically significant enhancement. This demonstrates that the proposed dual-teacher distillation strategy effectively strengthens the model's resilience to adversarial perturbations. More importantly, the model also exhibits an absolute gain of 3.5% in zero-shot accuracy, indicating that by learning more discriminative feature representations from the clean teacher, it acquires superior generalization capabilities rather than merely overfitting to adversarial examples. Furthermore, the increase in the overall Sum-acc metric (+2.48) further validates the comprehensive optimization effect of the proposed approach on the model's robustness and generalization performance.

5.3.2. High-intensity attack

As the attack intensity (ϵ) increases, the distribution difference between adversarial samples and clean samples intensifies, and the performance of all models declines as expected. Under this extreme setting, our method performs close to the baseline in terms of robustness, but maintains an advantage of approximately 1.6% in clean accuracy consistently. This indicates that our method does not lose robustness in extreme adversarial environments, and at the same time successfully enables the student model to learn representations that are closer to the essential features of natural images, thereby achieving better performance on clean data.

Table 1. Performance of the benchmark method and the proposed method under the MMT-ARD framework on ViT-B-32, ResNet-50, ResNet-101 and ViT-B-32-Lora models

Method	eps	CLIP ViT-B-32				CLIP RN50				CLIP RN101				CLIP ViT-B-32-Lora			
		acc	racc	sum-acc	clip-zero	acc	racc	sum-acc	clip-zero	acc	racc	sum-acc	clip-zero	acc	racc	sum-acc	clip-zero
baseline	1	61.84	49.00	110.84	26.40	43.92	23.92	67.84	6.5	45.84	20.44	66.28	3.8	43.24	22.46	65.70	16.10
	2	61.84	34.56	96.40	19.20	43.92	10.14	54.06	3.1	45.84	7.54	53.38	1.0	43.24	9.46	52.70	10.40
	3	61.84	22.72	84.56	14.00	43.92	4.04	47.96	1.8	45.84	2.26	48.1	0.5	43.24	2.74	45.98	5.0
	4	61.84	13.76	75.56	9.90	43.92	1.28	45.2	1.0	45.84	0.62	46.46	0.1	43.24	0.74	43.98	2.6
our	1	63.48	49.34	112.82	27.10	46.56	25.36	71.92	9.0	49.48	27.30	76.78	13.0	45.76	23.06	68.82	17.2
	2	63.48	34.78	98.26	19.60	46.56	10.94	57.5	5.1	49.48	12.46	61.94	6.4	45.76	9.24	55.0	7.5
	3	63.48	22.24	85.72	13.80	46.56	4.28	50.84	3.0	49.48	4.78	54.26	3.6	45.76	2.52	48.28	5.0
	4	63.48	12.92	76.42	9.60	46.56	1.46	48.02	1.6	49.48	1.62	51.10	2.0	45.76	0.62	46.38	2.8

5.3.3. Generalization verification

The results on the ResNet architecture further verify the universality of our method. For RN101, our method achieves absolute improvements of 3.64% in clean accuracy and 2.52% in robust accuracy. Most importantly, its robust accuracy has more than doubled (a relative increase of 111.5%), while the zero-shot performance has improved by 3.1% (a relative increase of 720%). These results demonstrate that the proposed dual-teacher distillation strategy is effective across models with varying capacities and architectures. In particular, it substantially enhances the adversarial robustness of classical architectures like ResNet while preserving strong transferability and generalization performance.

Taking the above analysis together, our multi-teacher distillation method can work effectively on multiple architectures such as ViT and ResNet, and its core advantages are reflected in: 1) significantly improving the robustness and generalization ability of the model under common low-intensity attacks; 2) Maintain competitiveness under high-intensity attacks and optimize the essential feature representation of the model; 3) It shows excellent generalization for different model architectures; 4) Perfect compatibility with efficient parameter fine-tuning technology, with high practical value. Figure 3 shows the visualization of the experimental results. Figure 3. (a) represents the original, clean input image,(b) represents the Grad-CAM heat map generated by the adversarial teacher (ViT-L-14) when processing the adversarial examples,(c) represents the Grad-CAM heat map generated by the clean teacher (ViT-L-14) when processing the clean original image, and (c) represents the Grad-CAM heat map generated by the clean teacher (ViT-L-14) when processing the clean original image. Heat map of (d) the original student model without distillation (ViT-B-32) on the original image,(e) the student model distilled by our proposed multi-teacher method on the original image, and (f) the student model distilled using only a single teacher (adversarial teacher). The figure clearly reveals different models (teacher vs. student) and different methods (baseline vs. student). Our approach) fundamental differences in the basis for decision making.

5.4. Ablation Study

To comprehensively analyze the performance of our proposed MMT-ARD framework and verify the effectiveness of the contribution of each component, we conducted systematic ablation studies. This section addresses three core questions: (1) What improvements are brought by introducing the clean teacher and its integration strategy? (2) How do different loss function designs affect the trade-off between accuracy and robustness of the model? (3) How should the supervisory signals from multiple teachers be balanced to achieve optimal performance? We explore these aspects through controlled experiments, isolating the effect of each factor.

5.4.1. Path-separated dual teacher strategy

This experiment evaluates the necessity of introducing a clean teacher and a confidence-based weighting strategy. We compared three strategies: 1. Baseline: Uses only the adversarial teacher model (ViT-L-14 PMG Fast2) with the student model ViT-B-32. 2. Average: Uses both the adversarial and clean teachers; their output embeddings are averaged with equal weights. 3. Path-Separated Dual Teachers (**Ours**): Employs both teachers, where their predictions are dynamically weighted and fused based on confidence levels.

As shown in Table 2, introducing a clean teacher consistently improves performance. Compared with the baseline, the equal-weight averaging strategy achieves minor improvements of +0.26% in clean accuracy and +0.16% in robust accuracy, with a Sum-acc increase of 0.42%. This demonstrates that discriminative features learned from the clean teacher (derived from natural image distributions) complement the robust features of the adversarial teacher.

However, our path-separated dual-teacher strategy further enhances performance. While maintaining robust accuracy (racc: 34.72%), the clean accuracy improves by 0.38% over the baseline. This indicates that allowing the clean teacher to focus on generating highly discriminative target embeddings for the original images provides a better learning target for the student model, thereby improving classifying

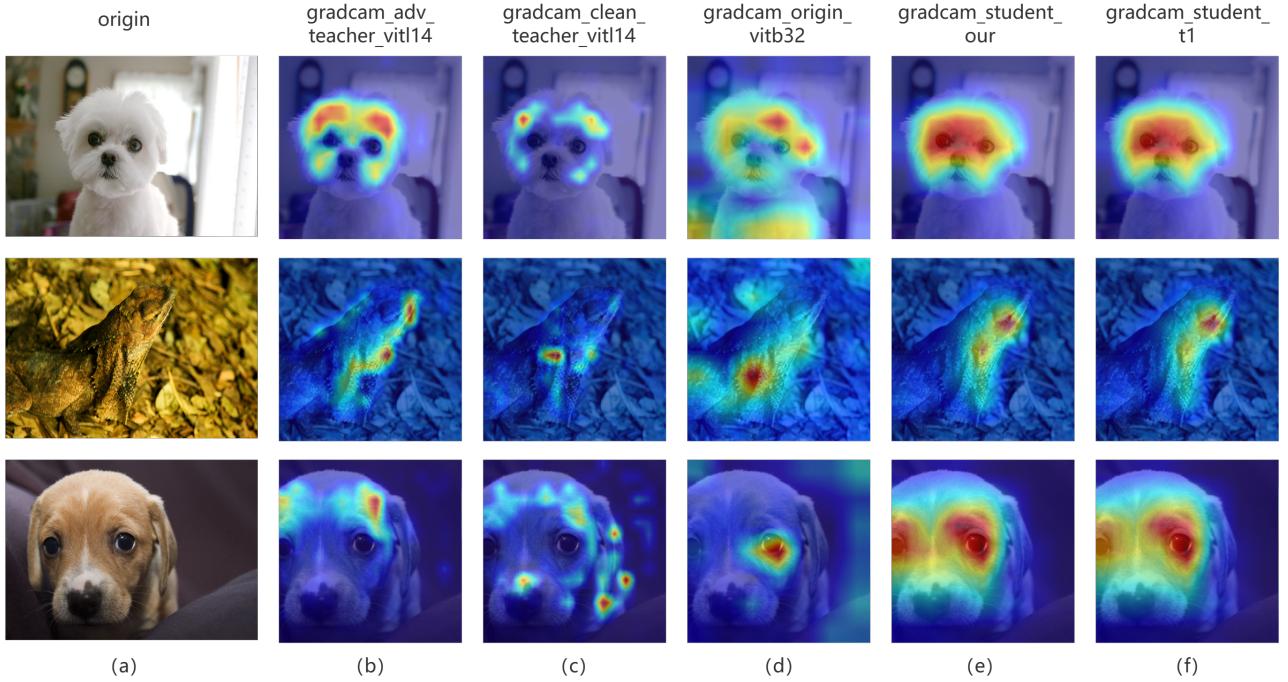


Figure 3. Heatmaps of the models for different teacher-student pairs.

fication performance. Compared with the simple Average strategy: the clean accuracy further improves by 0.12%, confirming that naive output fusion is suboptimal.

Table 2. Experimental results of different combinations of teachers. CA: Clean Accuracy, RA: Robust Accuracy, Baseline: (Adv. Teacher Only)

Strategy	CA (acc)	RA (racc)	Sum-acc
Baseline	61.96	34.56	96.52
Average	62.22	34.72	96.94
Weighted (Ours)	63.48	34.78	98.26

5.4.2. Dynamic weighting strategy based on teachers' confidence

This experiment compare three configurations: 1. Single-KL (Baseline): Uses only the adversarial teacher's output as soft labels to compute the KL divergence loss. 2. Dual-KL (0.5:0.5): Computes KL divergence from both teachers with equal (1:1) weighting. 3. Dual-KL + Adaptive Norm: Extends Dual-KL by introducing an adaptive normalization loss. As shown in Table 3, this experiment clearly illustrates the accuracy–robustness trade-off. When switching from Single-KL to fixed-weight Dual-KL loss, the model learns extremely discriminative features from the clean teacher, resulting in a sharp increase in clean accuracy by +10.18%. However, this aggressive optimization deviates significantly from the robust feature space guided

by the adversarial teacher, causing a sharp drop in robust accuracy by -23.02%. This indicates that giving equal weight to both teachers in the loss function leads to severe gradient conflicts in the optimization objective, making it difficult for the student model to simultaneously fit two highly divergent distributions. After adding the Adaptive Norm loss, the clean accuracy is further increased, but robustness is almost completely lost confirming that static fusion cannot effectively balance the competing learning signals. In contrast, incorporating a dynamic confidence-based weighting strategy significantly improves overall performance: compared to the baseline, clean accuracy is significantly improved by +1.52%, robust accuracy reaches 34.78%, and Sum-acc significantly gains +1.74%.

These results demonstrate that static averaging is sub-optimal, while dynamic weighting enables adaptive balancing. When the adversarial teacher exhibits high confidence, the model prioritizes its supervision to preserve robustness; when confidence is low, it relies more on the clean teacher's discriminative features to enhance accuracy. This adaptive cooperation between teachers is key to achieving balanced and superior performance.

5.4.3. Loss weight

Based on the findings in Section 5.4.2, we conducted an in-depth analysis to optimize the accuracy–robustness trade-off by fine-tuning the loss weight ratio between the two teachers (λ_{adv} : λ_{org}). The experiment successfully found

Table 3. Experimental results for different loss function designs.
CA: Clean Accuracy, RA: Robust Accuracy.

Loss Design	CA (acc)	RA (racc)
Single-KL (Baseline)	61.96	34.56
Dual-KL (0.5:0.5)	72.14	11.54
Dual-KL + Adaptive Norm	73.26	0.34

the optimal operation point (Sweet Spot). As shown in Table 4, when the weight ratio is set to 3:0.5 (i.e., $\lambda_{adv} / \lambda_{org} = 6$), the model achieves an optimal balance between clean accuracy (63.88%) and robust accuracy (34.42%). Both indicators at this point are significantly better than the Dual-KL (0.5:0.5) setting, and the robustness returned to a level comparable to the baseline, while the clean accuracy maintained an improvement of nearly 2%. These results highlight three key insights: 1. Adversarial supervision should dominate the training process: a higher λ_{adv} ratio is a prerequisite for maintaining model robustness. This is in line with the essence of adversarial training, that is, the model must prioritize learning stable decision boundaries. 2. Clean supervision refines representations: A small but non-zero λ_{org} weight is sufficient to provide the necessary discriminative signal, effectively refining the basic feature representation learned from adversarial training, thereby improving the clean accuracy without compromising its stability. 3. Balance is feasible: Through strict weight tuning, a new Pareto Optimal point can be found, breaking through the trade-off boundary between robustness and accuracy without significantly sacrificing robustness, and achieving an overall performance improvement.

Table 4. Experimental results of different loss weight proportions.CA: Clean Accuracy, RA: Robust Accuracy

Weight Ratio $\lambda_{adv} : \lambda_{org}$	CA (acc)	RA (racc)
1 : 0.5	71.26	16.18
2 : 0.5	68.72	21.84
3 : 0.5	63.88	34.42
3.5 : 0.5	63.74	34.44
3 : 1	69.88	18.76
7 : 0.3	62.88	34.60

5.4.4. Quantitative analysis of gradient

To quantitatively evaluate the consistency of visual attention regions between different distillation models and their teacher model, we adopt a method based on Grad-CAM feature map subtraction followed by L_2 norm computation. The numerical value intuitively reflects the degree of difference in the attention region between the models. The resulting L_2 norm intuitively reflects the degree of discrepancy between the attention regions: a smaller value indicates greater similarity between the student’s and teacher’s

gradient feature maps, implying stronger alignment with the teacher’s guidance.

As shown in Table 5, analyses conducted on three validation set images (ILSVRC2012_val_00004748, ILSVRC2012_val_00012820, ILSVRC2012_val_00014409) reveal that the proposed method consistently achieves significantly lower L_2 norm values than ViT-B-32, whether compared against the clean or adversarial teacher. This strongly proves that our approach can effectively make the student model learn and inherit the key feature attention regions of the teacher model, thereby improving feature representation transfer efficiency.

In the path distilled from adv.teacher, our method achieves L_2 values (2157, 2296, 2571) lower than or equal to the baseline method (2175, 2316, 2580) on all three images, indicating a slight but consistent advantage in capturing the attention mechanism of the robust teacher model. It reflects the positive effect of the introduced module. Therefore, from the perspective of gradient feature similarity, this experiment confirms that the multi-teacher distillation framework proposed in this paper can effectively promote the student model to align the visual attention of the teacher model more accurately, thus ensuring the effectiveness of knowledge distillation at the feature level, which lays a foundation for the performance improvement of the final model.

Table 5. Comparison of the knowledge distillation effectiveness of Clean teacher (Cle_T) and adversarial teacher (adv_T) models for ViT-B-32, baseline, and MMT-ARD (quantitative results on datasets of ILSVRC2012_val_00004748 (Val_1), ILSVRC2012_val_00012820 (Val_2), and ILSVRC2012_val_00014409 (Val_3) respectively.)

	Val_1	Val_2	Val_3
Cle_T to ViT-B-32	2613	2603	2655
Cle_T to Baseline	2104	2288	2630
Cle_T to ours	2107	2266	2598
adv_T to ViT-B-32	2650	2621	2779
adv_T to Baseline	2175	2316	2580
adv_T to ours	2157	2296	2571

6. Conclusion

This study have proposed a Multimodal Multi-teacher adversarial Robust distillation framework (MMT-ARD), which effectively solves the robustness problem of visual language models in adversarial environments through a dual-teacher knowledge fusion architecture and a dynamic weight allocation strategy. Experiments demonstrated that the proposed method improves robust accuracy of ViT-B-32 model by 4.32% and zero-shot accuracy by 3.5% on

the ImageNet dataset, and improves the training efficiency by 2.3 times. The results of this study provide new ideas and methods for the research on the adversarial robustness of multimodal models, and provide reliable technical support for artificial intelligence applications in safety-critical fields. Future work will focus on further optimizing the dynamic weight algorithm and extending the framework to more modalities and more complex application scenarios.

References

- [1] S Baluja and I Fischer. Adversarial transformation networks: Learning to generate adversarial examples. *arxiv* 2017. *arXiv preprint arXiv:1703.09387*. 1
- [2] Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song. Practical black-box attacks on deep neural networks using efficient query mechanisms. In *Proceedings of the European conference on computer vision (ECCV)*, pages 154–169, 2018. 2
- [3] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26, 2017. 1
- [4] Sizhe Chen, Zhengbao He, Chengjin Sun, Jie Yang, and Xiaolin Huang. Universal adversarial attack on attention and the resulting dataset damagenet. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2188–2197, 2020. 2
- [5] 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, 2009. 5
- [6] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4312–4321, 2019. 2
- [7] Inpyo Hong and Chang Choi. Knowledge distillation vulnerability of deit through cnn adversarial attack. *Neural Computing and Applications*, 37(12):7721–7731, 2025. 3
- [8] Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Beßongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4733–4742, 2019. 1
- [9] Yi Huang and Adams Wai-Kin Kong. Transferable adversarial attack based on integrated gradients. *arXiv preprint arXiv:2205.13152*, 2022. 2
- [10] Lin Li, Haoyan Guan, Jianing Qiu, and Michael Spratling. One prompt word is enough to boost adversarial robustness for pre-trained vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24408–24419, 2024. 2
- [11] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 2
- [12] Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. *arXiv preprint arXiv:2212.07016*, 2022. 2
- [13] Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. *arXiv preprint arXiv:2402.12336*, 2024. 1, 2
- [14] Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, pages 742–749, 2019. 1
- [15] Guoqiu Wang, Xingxing Wei, and Huanqian Yan. Improving adversarial transferability with spatial momentum. *arXiv preprint arXiv:2203.13479*, 2022. 2
- [16] Sibo Wang, Jie Zhang, Zheng Yuan, and Shiguang Shan. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24502–24511, 2024. 2
- [17] Xingxing Wei, Siyuan Liang, Ning Chen, and Xiaochun Cao. Transferable adversarial attacks for image and video object detection. *arXiv preprint arXiv:1811.12641*, 2018. 2
- [18] Tsung-Han Wu, Hung-Ting Su, Shang-Tse Chen, and Winston H Hsu. Revisiting semi-supervised adversarial robustness via noise-aware online robust distillation. *arXiv preprint arXiv:2409.12946*, 2024. 3
- [19] Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R Lyu, and Yu-Wing Tai. Boosting the transferability of adversarial samples via attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1161–1170, 2020. 1
- [20] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2730–2739, 2019. 2
- [21] Chuanguang Yang, Zhulin An, Linhang Cai, and Yongjun Xu. Hierarchical self-supervised augmented knowledge distillation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 1217–1223, 2021. 2
- [22] Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu, and Qian Zhang. Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12319–12328, 2022.
- [23] Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. Clip-kd: An empirical study of clip model distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15952–15962, 2024. 2
- [24] 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. [2](#)
- [25] Shiji Zhao, Jie Yu, Zhenlong Sun, Bo Zhang, and Xingxing Wei. Enhanced accuracy and robustness via multi-teacher adversarial distillation. In *European Conference on Computer Vision*, pages 585–602. Springer, 2022. [3](#)
- [26] Shiji Zhao, Ranjie Duan, Xizhe Wang, and Xingxing Wei. Improving adversarial robust fairness via anti-bias soft label distillation. *Advances in Neural Information Processing Systems*, 37:89125–89149, 2024. [3](#)
- [27] Shiji Zhao, Xizhe Wang, and Xingxing Wei. Mitigating accuracy-robustness trade-off via balanced multi-teacher adversarial distillation. *IEEE transactions on pattern analysis and machine intelligence*, 46(12):9338–9352, 2024. [3](#)
- [28] Bojia Zi, Shihao Zhao, Xingjun Ma, and Yu-Gang Jiang. Revisiting adversarial robustness distillation: Robust soft labels make student better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16443–16452, 2021. [2](#)