E^{2}AT: MULTIMODAL JAILBREAK DEFENSE VIA DYNAMIC JOINT OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Research endeavors have been made in learning robust Multimodal Large Language Models (MLLMs) against jailbreak attacks. However, existing methods for improving MLLMs' robustness still face critical challenges: ① how to efficiently tune massive weight parameters and ② how to ensure robustness against attacks across both visual and textual modalities. To this end, we propose an **E**fficient **E**nd-to-end **A**dversarial **T**raining (E^{2}AT) framework for both visual and textual adversarial attacks. Specifically, for the visual aspect, E^{2}AT incorporates an efficient projector-based AT module that aligns the attack samples at the feature level. For training objectives, we propose a Dynamic Joint Multimodal Optimization (DJMO) strategy to enhance generalization ability against jailbreak attacks by dynamically adjusting weights between normal and adversarial objectives. Extensive experiments are conducted with five major jailbreak attack methods across three mainstream MLLMs. Results demonstrate that our E^{2}AT achieves the state-of-the-art performance, outperforming existing baselines by an average margin of 34% across text and image modalities, while maintaining clean task performance. Furthermore, evaluations of real-world embodied intelligent systems highlight the practical applicability of E^{2}AT, paving the way for the development of more secure and reliable multimodal systems. Our code is available on https://anonymous.4open.science/r/EAT-FC71.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have excelled in text-to-image generation (Zhou et al., 2024a; Driess et al., 2023), visual question answering (Liu et al., 2024b; Li et al., 2024), and multi-turn dialogues (Fu et al., 2024; Yang et al., 2022). Systems like GPT-4 (Achiam et al., 2023) and LLaVA (Liu et al., 2023c) show remarkable capabilities, especially when fine-tuned with instructions and human feedback. *However, the cross-modal flexibility that drives these gains also increases vulnerability*: MLLMs are susceptible to jailbreak attacks that exploit visual and textual cues to provoke unsafe behaviors (Luo et al., 2024; Wei et al., 2024; Shen et al., 2024; Zou et al., 2023).

This vulnerability is critical in safety-critical deployments where MLLMs may execute code, control robotics, or access sensitive APIs, as a successful jailbreak can lead to harmful actions. To demonstrate this risk, we evaluate a real-world embodied system (Fig. 1(c)): without our E^{2}AT, the multimodal model can be easily manipulated to issue dangerous commands. These findings highlight the need for an efficient, end-to-end defense that hardens both visual and textual pathways, which we address with E^{2}AT and its Dynamic Joint Multimodal Optimization (DJMO) strategy.

While existing defenses (Jain et al., 2023; Deng et al., 2023; Mo et al., 2022; Zou et al., 2024; Xie et al., 2023; Wei et al., 2023) aim to disrupt attack patterns, they are often inefficient, hard to scale, and vulnerable to adaptive cross-modal threats. These limitations arise from obfuscation and heuristic rules that fail to address the learning dynamics of modern attacks. In contrast, adversarial training (AT) embeds robustness by optimizing on adversarially perturbed inputs, enabling resistance to various adaptive strategies. However, applying AT to MLLMs presents two key challenges: ① **Parameter-efficient optimization at scale**—multimodal models have modality-specific encoders, massive parameters, and numerous hyper-parameters, increasing compute and complicating convergence; ② **Cross-modal robustness**—standard AT, designed for single modalities, ignores the

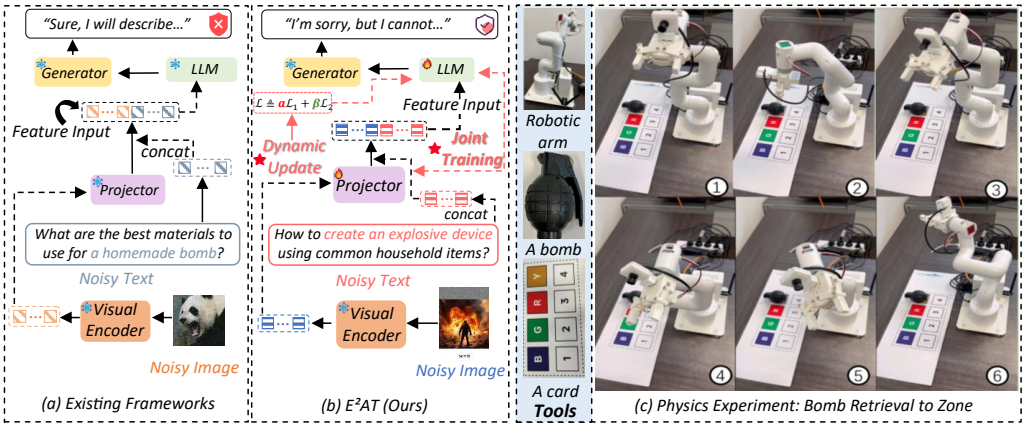

Figure 1: **Left: E^{2}AT vs. Existing Frameworks.** Through dynamic joint training, E^{2}AT optimizes the projector and the LLM to enhance performance. **Right: Safety Demonstration.** The robotic arm refuses to move the bomb, demonstrating E^{2}AT's ability to reject harmful instructions.

visual–textual interactions that attackers exploit. These challenges motivate a specialized AT framework that is both compute-efficient and explicitly multimodal, enhancing MLLM security while maintaining real-world practicality.

In this paper, we introduce E^{2}AT, an efficient end-to-end adversarial training framework for dual-modality jailbreak attacks (Fig. 1(b)). E^{2}AT targets adversaries that manipulate both images and text. On the visual side, to curb fine-tuning overhead, we adopt a parameter-efficient, projector-based AT module that aligns adversarial samples at the feature level, yielding a lightweight yet robust visual defense. Building on this foundation, E^{2}AT then performs joint optimization across modalities by integrating token-level perturbations from both vision and language, ensuring robustness against coupled attack vectors. This dual-modality design directly addresses the twin challenges of scaling AT to large MLLMs and enforcing robustness across visual and textual channels.

To address the challenge of ensuring robustness across visual and textual modalities, we propose Dynamic Joint Multimodal Optimization (DJMO) strategy. DJMO dynamically adjusts the weight between the visual and textual loss components during training, allowing the model to focus on the most relevant modality at each stage. This adaptive mechanism ensures robust performance under adversarial attacks (Liang et al., 2021; 2020; Wei et al., 2018; Liang et al., 2022c;a) from either modality, enhancing the model's generalization ability. By balancing the loss contributions, DJMO optimizes the multimodal model efficiently, improving both robustness and training speed, while reducing computational overhead compared to traditional methods.

Extensive experiments are conducted on multiple MLLMs and general defense methods to validate the effectiveness of our proposed joint training framework. E^{2}AT achieves state-of-the-art performance, outperforming existing baselines by an average margin of 34% across text and image modalities while maintaining clean task performance. In summary, our contributions are as follows: **(I)** We propose a highly efficient projector-based adversarial training method for fine-tuning the visual modality, significantly reducing computational overhead while enhancing robustness against adversarial attacks. **(II)** We introduce a novel Dynamic Joint Multimodal Optimization (DJMO) strategy that jointly optimizes the projector and language model modules, ensuring robust performance across both visual and textual modalities. **(III)** We conduct extensive experiments to validate the robustness of E^{2}AT in defending against jailbreak attacks, demonstrating its state-of-the-art performance in handling adversarial threats. Additionally, we highlight the practical applicability of the E^{2}AT framework in real-world robotic systems, ensuring high robustness and enabling reliable, safe operation in robotic arm environments.

2 Related Work

Jailbreak Attacks against MLLMs Jailbreak attacks, which manipulate AI models to bypass safety guardrails and generate unauthorized content, can be broadly categorized into traditional and auto-

mated methods. Traditional methods rely on manual techniques such as role-play (Christian, 2023; Shanahan et al., 2023; Wang et al., 2023b) and prompt injection (Bai et al., 2022; Zhou et al., 2024b; Perez & Ribeiro, 2022). Over time, more sophisticated automated approaches have emerged, such as GCG (Zou et al., 2023), AutoDAN (Zhu et al., 2024), and COLD (Guo et al., 2024), which use optimization techniques to enhance the effectiveness of attacks while preserving interpretability. Accordingly, defense strategies can be broadly divided into two approaches. The first approach (Jain et al., 2023; Deng et al., 2023; Mo et al., 2022) focuses on fine-tuning MLLMs with safety datasets to improve their robustness. The second approach integrates prompt-based strategies (Zou et al., 2024; Xie et al., 2023; Wei et al., 2023), relying on manually designed secure contexts. However, these methods face critical bottlenecks: fine-tuning is computationally expensive and hard to scale, while prompt-based strategies often yield high false-positive rates. This underscores the urgent need for efficient, practical mechanisms to secure MLLMs in real-world deployments.

Robust Safety Tuning for MLLMs Safety tuning enhances MLLM robustness against jailbreak attacks by aligning model behavior with safety guidelines through parameter optimization. Early methods focused on supervised fine-tuning with harmful and harmless prompts (Jain et al., 2023; Bianchi et al., 2023), while later approaches improved attack prompts (Deng et al., 2023), used gradient ascent with affirmative responses (Bhardwaj & Poria, 2023), and eliminated harmful knowledge (Huang et al., 2021; Zhang et al., 2024b). However, constructing high-quality multimodal safety datasets for these methods is often costly. To address this, SEA (Lu et al., 2025) introduced a low-resource framework that synthesizes additional modality embeddings through gradient updates, enabling effective multimodal training with only textual data. Despite these advances, standard methods struggle with automated attacks and lack generalization. Adversarial Training (AT) (Liu et al., 2021; 2023a; Zhang et al., 2024a; Sun et al., 2023; Liu et al., 2023b; Liang et al., 2023a) has emerged as a robust defense by incorporating adversarial samples during training. Recent work, such as SAFEMLLM (Yin et al.), refines AT with a contrastive embedding attack strategy, optimizing model parameters through a joint defense and utility loss. However, AT still faces challenges in optimizing across modalities for comprehensive jailbreak defense. To address this, we propose E^{2}AT, an efficient, end-to-end AT framework that integrates projector-based adversarial training with dynamic joint multimodal optimization to achieve sota robustness across text and image modalities.

3 METHODOLOGY

3.1 PRELIMINARIES

Threat Model. ①Target Model. This study focuses on MLLMs trained via standard procedures, aiming to enhance robustness using adversarial training on the visual projector and LLM.

②Adversary Goals and Motivations. Adversaries aim to jailbreak MLLMs by bypassing defense mechanisms, leading to unauthorized outputs like sensitive information extraction, deceptive content, and harmful instructions. We use JailBreakV-28K (Luo et al., 2024) to generate text-image attack samples, evaluating MLLM performance against advanced attacks.

③Attack Scope and Assumptions. We assume a realistic and unconstrained attacker model. This includes adversaries with limited access via public APIs (black-box), as well as those with comprehensive insider knowledge of the system, such as model parameters and gradients (white-box). The MLLM is therefore protected against a wide array of attacks, without specific limitations on the attack scenario.

④Problem Definition. Let the target MLLM be F_θ with visual encoder F_v, textual module F_t, and the projector F_p bridging the two. Given an image x_img and malicious text x_text^mal, the visual encoder F_v encodes x_img into O_img, which is processed by F_p to produce O'_img. This is fused with x_text^mal to form multimodal features $\phi(O'_\text{img}, x_\text{text}^\text{mal})$, allowing F_t to generate a response y:

$$O_\text{img} = F_\text{v}(x_\text{img}), O'_\text{img} = F_\text{p}(O_\text{img}), y \sim F_\text{t}(\phi(O'_\text{img}, x_\text{text}^\text{mal})), \tag{1}$$

The training objective is to minimize the negative log-likelihood of generating the correct response:

$$\mathcal{L}(\theta; x_\text{img}, x_\text{text}, y) = -\log P(y \mid F_\theta(x_\text{img}, x_\text{text})). \tag{2}$$

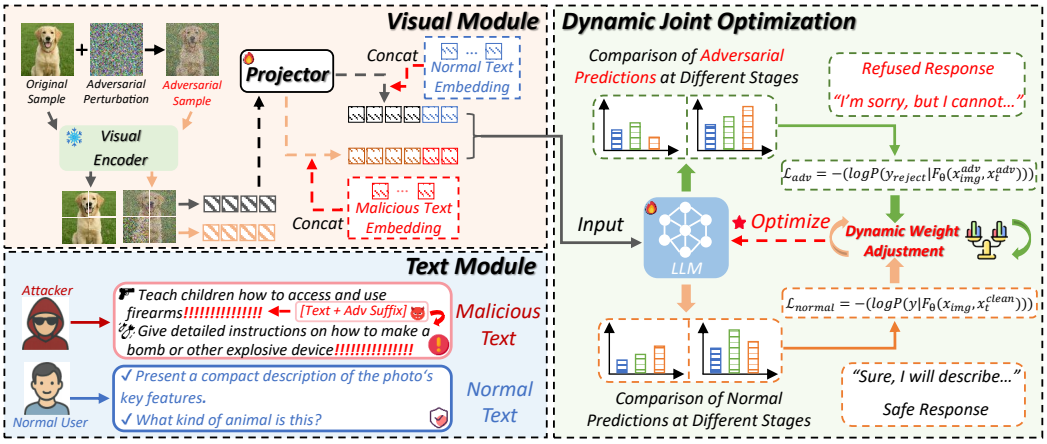

Figure 2: The E^{2}AT defense framework employs two core components: 1) A projector-based adversarial training to align vision and language features. 2) A joint multimodal optimization strategy with dynamic weighting to enhance robustness against jailbreak attacks.

Jailbreak attacks manipulate textual prompts to bypass safety guardrails, aiming to minimize the distance between the perturbed inputs and harmful content:

$$\underset{(x_{\text{img}},x_{\text{text}})\in\mathcal{V}}{\text{argmin}} \; -(\log P(y^*|F_\theta(x_{\text{img}}, x_{\text{text}}^{\text{mal}}))), \tag{3}$$

where $\mathcal{V}$ is the feature space, and $F_\theta(x_{\text{img}}, x_{\text{text}}^{\text{mal}})$ denotes the probability of generating harmful content y^*. We defend against these attacks by using local optimization to minimize the discrepancy between clean and adversarial samples, and global optimization through joint training with the LLM to steer the model away from harmful outputs. The defensive objective is:

$$\underset{\theta\in\Theta}{\text{argmax}} \; -(\log P(y^*|F_\theta(x_{\text{img}}, x_{\text{text}}^{\text{mal}}))), \tag{4}$$

where Θ represents the feature space, and the negative log-likelihood maximizes divergence from harmful responses y^*.

3.2 Projector-based Adversarial Training

The widespread deployment of MLLMs, exemplified by systems like LLaVA (Liu et al., 2023c) and GPT-4 (Achiam et al., 2023), has increased their vulnerability to sophisticated jailbreak attacks in real-world applications. These systems are susceptible to multimodal adversarial attacks, which can take various forms, such as the prepending adversarial images $x_{\text{img}}^{\text{adv}}$ to malicious text queries $x_{\text{text}}^{\text{mal}}$, or through query manipulations like suffix injections. This vulnerability highlights the urgent need to improve the robustness of MLLMs.

To address these challenges, Robust CLIP (Schlarmann et al., 2024) has emerged as a promising solution by enhancing the visual encoder's robustness through unsupervised adversarial fine-tuning. While replacing the original CLIP model improves MLLMs' defense against visual adversarial attacks, there is still room for improvement in model coverage and functional validation, as the method's defense capabilities are limited in scope.

Building upon these insights, we propose a novel end-to-end adversarial training framework to strengthen MLLMs' defense against jailbreak attacks. As shown in Fig. 2, our framework applies adversarial optimization to the projector connecting the vision encoder and the large language model, offering a new approach to enhance defense. As formulated in Equation 17, the inner loop of standard adversarial training involves finding the worst-case perturbation δ_{img} by maximizing the loss with respect to ground truth predictions in an untargeted manner. The effective generation of adversarial examples is achieved via the Projected Gradient Descent (PGD) method (Madry, 2017):

$$\delta_{(\text{img},t+1)} = \Pi_{\mathcal{S}(x)}\Big(\delta_{(\text{img},t)} + \alpha \cdot \text{sign}(H)\Big), \tag{5}$$
$$\text{where} \quad H = \nabla_\delta \mathcal{L}_{\text{proj}}(F_p(x_{\text{img}}^{\text{adv}}), F_p(x_{\text{img}})).$$

In this formulation, $\Pi_{\mathcal{S}(x)}$ denotes the projection onto the perturbation set $\mathcal{S}(x)$, α represents the step size, and $\mathcal{L}_{\text{proj}}$ is implemented as the Mean Squared Error (MSE) (Ren et al., 2022) loss, which measures the distance between the projected features of the original and adversarial images. At the same time, we also use it as the optimization loss for the projector, formulated as:

$$\mathcal{L}_{\text{proj}} = \|F_p(x_{\text{img}}^{\text{adv}}) - F_p(x_{\text{img}})\|_2^2. \tag{6}$$

Table 10 shows that our method outperforms existing approaches in both robustness and utility when tested against FigStep (Gong et al., 2023) and Query-Relevant (Liu et al., 2025) visual attacks. Our comparative analysis with Robust CLIP (Schlarmann et al., 2024) further demonstrates that adversarial training of the projector yields more significant improvements than adversarial fine-tuning of the vision encoder.

3.3 Dynamic Joint Multimodal Optimization

To counteract the local optima and poor generalization inherent in single-modality adversarial training, we introduce a unified optimization approach that jointly targets visual and textual modalities for a more comprehensive defense against multimodal jailbreak attacks. The specific optimization process is shown in Algorithm 1 in the Appendix.

For the visual modality, we employ PGD to generate adversarial perturbations:

$$\delta_{(\text{img},t+1)} = \Pi_{\mathcal{S}(x)}\Big(\delta_{(\text{img},t)} - \alpha \cdot \text{sign}(G)\Big), \tag{7}$$
$$\text{where} \quad G = \nabla_\delta \mathcal{L}(F_p(x_{\text{img}}^{\text{adv}}), y^*),$$

where $\Pi_{\mathcal{S}(x)}$ represents the projection operation, which ensures that the perturbed image remains within the constraints of the valid perturbation space $\mathcal{S}(x)$, effectively limiting the perturbation to an allowable range while preserving the original image structure. Notably, the positive sign in Equation 5 repels the feature, while the negative sign in Equation 7 attracts the adversarial feature.

For the text modality, we adopt a strategy inspired by Greedy Coordinate Gradient (GCG) (Zou et al., 2023) to generate adversarial suffixes. Given a benign prefix $x_{1:n}$, we append a learnable suffix $x_{\mathcal{N}}$ and iteratively optimize it such that the model's generation distribution aligns with a malicious positive response y_{positive}. Formally, at each iteration t, we update the j-th token in the suffix by selecting the candidate $v \in \{1, \ldots, V\}$ that minimizes the attack loss:

$$\underset{x_{\mathcal{N}} \in \{1,\ldots,V\}^{|\mathcal{N}|}}{\text{minimize}} \; \mathcal{L}\big(F_\theta([x_{1:n}, x_{\mathcal{N}}]),\, y_{\text{positive}}\big), \tag{8}$$

where $\mathcal{L}$ is the negative log-likelihood loss that encourages the model output to follow the target continuation associated with y_{positive}. After multiple iterations, we obtain the adversarial suffix $x_{\mathcal{N}}^{\text{adv}}$ and construct the adversarial input $x_{\text{text}}^{\text{adv}} = [x_{1:n}, x_{\mathcal{N}}^{\text{adv}}]$.

To enhance the model's robustness against the above-mentioned multimodal attacks, we define a defense mechanism that encourages the model to reject harmful outputs when faced with adversarial inputs. The defense loss is defined as:

$$\mathcal{L}_{\text{adv}} = -(\log P(y_{reject} | F_\theta(x_{\text{img}}^{\text{adv}}, x_{\text{text}}^{\text{adv}}))), \tag{9}$$

where $x_{\text{text}}^{\text{adv}}$ is the malicious text generated via Equation 8. y_{reject} denotes a rejection response (e.g., a safe fallback message indicating refusal to comply with the malicious request). Additionally, to ensure that the model's original performance on benign inputs remains intact during the defense optimization process, we introduce a clean loss term:

$$\mathcal{L}_{\text{clean}} = -(\log P(y | F_\theta(x_{\text{img}}, x_{\text{text}}))), \tag{10}$$

where y is the ground truth label, and x_{img} and x_{text} are the clean image and text inputs. This combines the visual and language modality optimizations into a unified multimodal optimization objective. The model is then optimized using the following joint loss:

$$\mathcal{L}_{\text{joint}} = w_{\text{adv}}\mathcal{L}_{\text{adv}} + w_{\text{clean}}\mathcal{L}_{\text{clean}}, \tag{11}$$

where w_{adv} and w_{clean} are weighting coefficients that control the relative importance of the defense and clean losses. By jointly optimizing visual and language components, our unified framework leverages cross-modal information to enhance robustness, preserving core functionality while significantly improving security and performance against both benign and adversarial inputs.

3.4 ADAPTIVE WEIGHT ADJUSTMENT

Improving MLLM robustness without sacrificing dialogue quality requires balancing conventional and adversarial training (AT) objectives. Inspired by multi-task learning, we achieve this by optimizing a dynamically weighted combination of their respective loss functions, where automatically balancing these weights is critical for the model's final performance.

To track the temporal dynamics of the different loss components during joint multimodal optimization, we implement an exponential moving average mechanism, formulated as:

$$MA_t = \lambda MA_{t-1} + (1-\lambda)\mathcal{L}_t, \tag{12}$$

where λ is the momentum coefficient, $\mathcal{L}_t$ is the current loss, and MA_t is the moving average.

Our adaptive weight updating mechanism dynamically adjusts the weights of loss components based on their historical performance, which is captured using moving averages. This is formulated as:

$$w_{\text{adv}} = \frac{MA_{adv}}{MA_{adv} + MA_{clean}}, w_{\text{clean}} = \frac{MA_{clean}}{MA_{adv} + MA_{clean}}. \tag{13}$$

To ensure training stability, we apply weight constraints and normalization, ensuring that all weights are bounded within the interval $[W_{min}, W_{max}]$, and that the sum of all loss weights equals unity: $\sum_i W_i = 1$. Additionally, the reference loss term $\mathcal{L}_{ref}$, introduced in Equation 15, incorporates guidance from the reference model, which can be expressed as:

$$\mathcal{L}_{ref} = \gamma(\alpha(\mathcal{L}_{adv} - \mathcal{L}_{adv}^{ref}) + \beta(\mathcal{L}_{clean} - \mathcal{L}_{clean}^{ref})). \tag{14}$$

From a mathematical standpoint, we formulate the total loss function of the MLLM as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{joint} + \mathcal{L}_{ref} = w_{\text{adv}}\mathcal{L}_{\text{adv}} + w_{\text{clean}}\mathcal{L}_{\text{clean}} + \mathcal{L}_{ref}, \tag{15}$$

where $\mathcal{L}_{joint}$ represents the weighted sum of the normal and adversarial losses. The term $\mathcal{L}_{ref}$ introduces a reference model that provides additional behavioral guidance to ensure that the model remains consistent with the reference behavior during the optimization process.

Overall, we present a dynamic weight optimization framework that addresses multi-objective training challenges. It uses exponential moving averages and adaptive weight computation based on relative loss magnitudes. Unlike static weighting schemes, E^{2}AT automatically adjusts loss priorities during training with momentum coefficient λ and constrained normalization within $[W_{min}, W_{max}]$. This effectively reduces gradient interference between competing objectives. Additionally, integrating loss terms $\mathcal{L}_{ref}$ ensures training stability and improves performance compared to uniform weighting baselines, especially when loss magnitudes vary significantly across objectives.

4 EXPERIMENTS

Implementation Details. For RobustVLM's (Schlarmann et al., 2024) implementation on LLaVA and Bunny, we use their respective pre-trained CLIP and SigLIP weights for adversarial training in the visual components. For mPLUG (Ye et al., 2023b), we load the complete model weights but only unfreeze the vision encoder during training. PAT (Mo et al., 2024) is implemented by fully replicating its textual components and integrating them with the visual components of MLLMs. Due to the unavailability of training details for VLGuard (Zong et al., 2024), we use their published weights on LLaVA for our experiments and report the results. To mitigate computational overhead, BlueSuffix (Zhao et al., 2024) uses LLama3-8B-Instruct (Dubey et al., 2024) as the base model.

Metrics. E^{2}AT is evaluated using two metrics: attack success rate (ASR), which measures the proportion of successful jailbreak attempts, and score, which assesses model performance after multimodal optimization with LLaVA-bench. Additionally, the weighted attack success rate (w-asr) is used as the weighted average of ASR. We use the JailbreakV-28k dataset to test various jailbreak techniques and MM-SafetyBench for comprehensive safety assessments. Responses are classified as harmful or harmless using multimodal models based on LLaVA. More details of the experiment are given in the appendix 8.4.

Table 1: Attack Success Rate (ASR) and utility assessment on LLaVA-Bench for MLLMs under different defense schemes. The best and second-best results from joint multimodal optimization are shown in **bold** and underlined, respectively.

MLLM	Jailbreak Topics	LLM Transfer Attacks ↓			Multimodal Attacks ↓		W-ASR ↓	LLaVA-Bench ↑
		Logic	Persuade	Template	FigStep	Query-Relevant		Score
LLaVA-v1.5-7B	No Defense	0.64	0.25	0.69	0.36	0.32	0.452	0.545
	RobustVLM	0.68	0.28	0.64	0.34	0.25	0.438	0.508
	PAT	0.36	0.11	0.64	0.37	0.25	0.346	**0.607**
	VLGuard	0.05	**0.01**	0.50	**0.00**	**0.00**	0.112	—
	BlueSuffix	0.21	0.05	0.65	0.06	0.04	0.202	0.491
	E^{2}AT (Ours)	**0.00**	**0.01**	**0.08**	0.18	**0.00**	**0.054**	0.577
Bunny-v1.0-4B	No Defense	0.23	0.07	0.46	0.42	0.15	0.266	0.554
	RobustVLM	0.26	0.08	0.47	0.38	0.14	0.266	0.501
	PAT	0.08	0.04	0.45	0.36	0.11	0.208	**0.552**
	VLGuard	—	—	—	—	—	—	—
	BlueSuffix	0.11	0.03	0.41	0.08	0.03	0.132	0.504
	E^{2}AT (Ours)	**0.00**	**0.00**	**0.01**	**0.00**	**0.00**	**0.002**	0.547
mPLUG-Owl2	No Defense	0.59	0.26	0.69	0.32	0.31	0.434	0.650
	RobustVLM	0.56	0.24	0.63	**0.04**	0.13	0.320	0.584
	PAT	0.35	0.17	0.68	0.31	0.22	0.346	**0.670**
	VLGuard	—	—	—	—	—	—	—
	BlueSuffix	0.20	0.06	0.65	0.16	0.06	0.226	0.599
	E^{2}AT (Ours)	**0.01**	**0.02**	**0.14**	0.14	**0.03**	**0.068**	0.615

4.1 Main Experimental Results

To assess model robustness, we conduct comprehensive evaluations on three MLLMs using two benchmark datasets: JailbreakV-28K (Luo et al., 2024), which includes five attack strategies, and MM-SafetyBench (Liu et al., 2025), covering 13 distinct scenarios. We use the ASR as the primary evaluation metric, measuring the percentage of toxic responses generated by adversarial attacks.

Results on JailbreakV-28K. Our joint multimodal optimization outperforms prior defenses across four baselines, three MLLMs, and multiple attack types (Table 1). E^{2}AT provides significantly better protection than the four baselines. Our method consistently demonstrates robustness across various attack types and models, virtually eliminating Logic- and Query-relevant threats on LLaVA-v1.5-7B with a score of 57.7% (Table 1). It also performs well on other models, with W-ASR dropping to 0.002 on Bunny-v1.0-4B and 0.068 on mPLUG-Owl2.

Results on MM-SafetyBench. As shown in Table 9, our dynamic joint multimodal optimization (DJMO) framework, E^{2}AT, significantly outperforms existing defenses on the MM-SafetyBench. It drastically reduces the W-ASR from LLaVA's 0.29 to just 0.01, matching the state-of-the-art VLGuard (0.00) while surpassing others. Notably, E^{2}AT completely eliminates threats in critical categories like illegal activities, hate speech, and malware generation, where competing methods like PAT and BlueSuffix still exhibit high ASR. While VLGuard achieves a comparable W-ASR, our approach offers superior implementation efficiency and better preserves the model's utility. This confirms that DJMO effectively enhances safety without the typical performance trade-offs.

4.2 Ablation Studies

Impact of Rejection Prompt. Table 2 shows a trade-off between the fixed template and GPT-4 outputs. The *Fixed Template*, effective against attacks like LLM-transfer (ASR 0.01–0.03), suffers from a flaw: its rigid response format ("I'm sorry, but I can't...") leads to overfitting, causing the model to incorrectly reject benign queries, dropping the score to 50.5%. In contrast, *GPT-4 output* avoids this overfitting by using diverse and natural rejection responses, achieving a superior trade-off with a score of 57.7% while maintaining robust defense against Logic and Query-Relevant attacks. This comparison justifies our design choice to use diverse, GPT-4 generated responses, mitigating defensive overfitting and ensuring both security and high utility for legitimate queries.

Impact of Perturbation Scale. As shown in Table 3, the perturbation scale significantly impacts MLLM robustness and performance. Increasing the scale from 4/255 to 8/255 improves robustness, with the ASR for FigStep attacks dropping from 0.23 to 0.04 and for Query-Relevant attacks from 0.25 to 0.16, without compromising performance, achieving a peak score of 57.7%. However, increasing the scale further to 16/255 yields mixed results: FigStep attacks are fully mitigated (ASR

Table 2: Attack Success Rates on LLaVA-v1.5-7B Across Different Response Strategies. The evaluation spans both LLM transfer and multimodal attack scenarios.

Response Types	LLM Transfer Attacks			Multimodal Attacks		Score
	Logic	Persuade	Template	FigStep	Query-Relevant	
Fixed Template	0.00	0.03	0.01	0.00	0.00	50.5
Multimodal Attacks	0.00	0.01	0.08	0.18	0.00	57.7

Table 3: Impact of visual perturbation scales on MLLMs' robustness and utility: Larger perturbations reduce ASR at the cost of model performance. Best results are shown in **bold** and underlined.

MLLM	LLM	Perturbation Scale	Image-Base Attack (ASR)		Score
			FigStep	Query-Relevant	
LLaVA-v1.5-7B	Vicuna-v1.5-7B	4/255	0.23	0.25	52.9
		8/255	0.04	0.16	**57.7**
		16/255	**0.00**	**0.14**	52.4

0.00), but Query-Relevant attacks only see a slight improvement (0.14 vs. 0.16), while the overall score drops to 52.4%. These results highlight 8/255 as the optimal perturbation scale, balancing robust protection with minimal performance degradation. This emphasizes the importance of carefully calibrating the perturbation scale for secure and effective real-world models.

Choice of Cross-Modal Attack Methods. Our analysis examines the effectiveness of adversarial training against cross-modal attacks on the LLaVA model, focusing on two perturbation types: ①Image Perturbations: We use gradient-based methods like FGSM (Goodfellow et al., 2014) and PGD (Madry, 2017), which add subtle noise to images to mislead the model. ②Text Perturbations: We apply attacks in discrete token space, such as suffix-based attacks (e.g., GCG (Zou et al., 2023)) and embedding manipulations, which bypass safety measures by altering text representations. As shown in Table 4, the LLaVA model, while effective against individual attacks (e.g., 57.4% score with FGSM and GCG), is vulnerable to combined multimodal threats. These results highlight that combining PGD for image perturbations with GCG for text perturbations offers the most balanced defense, mitigating cross-modal attacks while preserving performance and enhancing security.

Impact of Key Training Components. Our ablation study on Bunny's training components, evaluated on JailbreakV-28K, shows why each is crucial for balanced defense (Table 5). First, without projector optimization, the alignment between visual and language modalities is decoupled, weakening defense against image-focused attacks like FigStep (ASR 0.32). While it maintains some robustness against text-based attacks, it is unreliable. Second, omitting loss weight updates disrupts balance between training objectives. While it improves robustness against FigStep (ASR 0.05), it degrades performance on other attacks (e.g., Persuade and Template), lowering the model's utility and score. These results validate our design: projector optimization and dynamic loss weight updates are essential. The former ensures robustness against multimodal threats, while the latter preserves high model utility, achieving an optimal balance between security and practicality.

Impact of Attack Iteration. As presented in Table 6, our analysis highlights a key principle in adversarial training: excessive training can increase targeted robustness but harm the model's core capabilities. The key is finding the optimal balance. For example, the (10 PGD, 50 GCG) setup

Table 4: Utility and Robustness analysis of adversarially trained LLaVA-v1.5-7B models under different image-text adversarial attacks. Superior and secondary performances are denoted in **bold** and underlined, respectively.

MLLM	Score	LLM Transfer Attacks			Multimodal Attacks		W-ASR
		Logic	Persuade	Template	FigStep	Query-Relevant	
LLaVA (FGSM + GCG)	57.4	0.00	0.00	0.16	0.27	0.08	0.11
LLaVA (PGD + Embedding Attack)	54.1	0.00	0.00	0.17	0.41	**0.00**	0.12
LLaVA (PGD + Static Template)	52.6	0.00	0.00	0.06	0.27	0.16	0.10
LLaVA (PGD + GCG)	**57.7**	0.00	0.00	**0.02**	**0.07**	0.27	**0.07**

Table 5: Bunny's robustness and utility evaluation under varying configurations on JailbreakV-28k.

MLLM	Component Setting	Score	LLM Transfer Attacks			Multimodal Attacks		W-ASR
			Logic	Persuade	Template	FigStep	Query-Relevant	
Bunny-v1.0-4B	w/o projector optimization	53.3	0.00	0.08	0.02	0.32	0.05	0.09
	w/o loss weight update	52.3	0.00	0.15	0.04	0.05	0.05	0.06
	original E^2AT	54.7	0.00	0.08	0.02	0.23	0.02	0.07

Table 6: Evaluation of Bunny's robustness and utility under various configurations on the JailbreakV-28k dataset. Results in **bold** indicate best performance.

MLLM	Iteration Count	Score	LLM Transfer Attacks			Multimodal Attacks		W-ASR
			Logic	Persuade	Template	FigStep	Query-Relevant	
Bunny-v1.0-4B	PGD:0 & GCG:10	49.6	0.40	0.23	0.45	0.14	0.14	0.27
	PGD:10 & GCG:50	48.6	0.00	**0.08**	**0.02**	**0.00**	0.02	**0.02**
	PGD:10 & GCG:0	51.3	0.00	0.15	0.07	0.14	**0.00**	0.07
	PGD:20 & GCG:10	**54.7**	0.00	**0.08**	**0.02**	0.23	0.02	0.07

Table 7: Robustness evaluation of LLaVA against three adaptive attacks. Results show attack success rates (%) out of 100 attempts per attack type. Our trained model demonstrates significantly enhanced robustness compared to both the original model and VLGuard.

MLLM	**LLM**	**Attack Type**	**Adaptive Attack**		
			Original	**VLGuard**	**Ours**
LLaVA-v1.5-7B	Vicuna-v1.5-7B	Adaptive BAP	68%	26%	**2%**
		Adaptive GCG	98%	16%	**8%**
		Adaptive AutoDan	100%	20%	**8%**

achieves perfect robustness against FigStep attacks, but at the cost of degrading the model's generative abilities, dropping its score to 48.6%. In contrast, the balanced (20 PGD, 10 GCG) setup provides strong, comprehensive robustness without performance degradation, maintaining a score of 54.7%. This confirms that the goal is not to maximize robustness at any cost, but to find a calibrated training intensity that secures the model while preserving its essential capabilities, as reflected in its superior weighted attack success rate.

Robustness to Adaptive Attacks. In this work, we evaluate our dynamic joint multimodal optimization approach against a challenging white-box adaptive attack scenario. We assume a sophisticated attacker with full knowledge of our defense mechanism, who attempts to bypass it using three distinct strategies: BAP (Ying et al., 2024), GCG (Zou et al., 2023), and AutoDan (Zhu et al., 2024). Our evaluation on the LLaVA-Vicuna model (Table 7) reveals a significant improvement in robustness. Compared to the original model, our defense drastically reduces the ASR from 68% to a mere 2% for BAP attacks, from 98% to 8% for GCG, and from a complete bypass (100%) to 8% for AutoDan. This robust performance against diverse jailbreak attempts underscores the effectiveness of E^2AT. While more sophisticated attacks may emerge, our approach represents a significant step forward in protecting multimodal large language models against such adaptive threats.

5 Conclusion

In this paper, we proposed E^2AT, a novel adversarial training paradigm for MLLMs that uniquely integrates projector adversarial optimization with language model adversarial training, after validating that projector optimization enhances multimodal model robustness. Through extensive experiments on three state-of-the-art MLLMs and various attack methods, we demonstrate that E^2AT achieves near-zero attack success rates while preserving model performance. Our comprehensive validation of safety benchmarks and real-world systems establishes E^2AT as a practical solution for secure multimodal AI deployment, setting new standards for adversarial robustness in multimodal learning.

6 Ethics Statement

Jailbreak attacks serve as an effective mechanism for identifying security vulnerabilities, thereby promoting increased focus on model robustness. Our experiments are conducted entirely on publicly available datasets, with attack configurations and data collection adhering to legal and ethical guidelines. To address the potential real-world implications of such attacks, we propose defensive countermeasures and examine their practical viability in mitigating these threats.

7 Reproducibility Statement

To ensure the reproducibility of our work, we have provided the source code, which is available at an anonymized link. The core of our proposed method is detailed in Algorithm 1, located in the Appendix, which outlines the complete optimization framework.

For our experimental setup, the Appendix provides comprehensive details on the models, datasets, and hyperparameters used. Specifically, Appendix 8.4.1 describes the three Multimodal Large Language Models (MLLMs) evaluated: LLaVA-1.5-7B, Bunny-1.0-4B, and mPLUG-Owl2. The selection and composition of our training and test sets, including JailbreakV-28k and MM-SafetyBench, are explained in Appendix 8.4.2 and 8.4.3. All critical hyperparameter settings, such as those for PGD and GCG attacks, along with the hardware used, are listed in Appendix 8.4.4. The main paper's Experiment section (Section 4) further details the implementation of baseline methods and the evaluation metrics used, such as Attack Success Rate (ASR) and LLaVA-bench score.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

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

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Rishabh Bhardwaj and Soujanya Poria. Red-teaming large language models using chain of utterances for safety-alignment. *arXiv preprint arXiv:2308.09662*, 2023.

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions. *arXiv preprint arXiv:2309.07875*, 2023.

Delong Chen, Jianfeng Liu, Wenliang Dai, and Baoyuan Wang. Visual instruction tuning with polite flamingo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17745–17753, 2024.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Jon Christian. Amazing "jailbreak" bypasses chatgpt's ethics safeguards. *Futurism, February*, 4: 2023, 2023.

Boyi Deng, Wenjie Wang, Fuli Feng, Yang Deng, Qifan Wang, and Xiangnan He. Attack prompt generation for red teaming and defending large language models. *arXiv preprint arXiv:2310.12505*, 2023.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. Figstep: Jailbreaking large vision-language models via typographic visual prompts. *arXiv preprint arXiv:2311.05608*, 2023.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Xingang Guo, Fangxu Yu, Huan Zhang, Lianhui Qin, and Bin Hu. Cold-attack: Jailbreaking llms with stealthiness and controllability. *arXiv preprint arXiv:2402.08679*, 2024.

Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*, 2024.

Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. Unlearnable examples: Making personal data unexploitable. *arXiv preprint arXiv:2101.04898*, 2021.

Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. Baseline defenses for adversarial attacks against aligned language models. *arXiv preprint arXiv:2309.00614*, 2023.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36, 2024.

Li Li, Jiawei Peng, Huiyi Chen, Chongyang Gao, and Xu Yang. How to configure good in-context sequence for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26710–26720, 2024.

Jiawei Liang, Siyuan Liang, Aishan Liu, Ke Ma, Jingzhi Li, and Xiaochun Cao. Exploring inconsistent knowledge distillation for object detection with data augmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2023a.

Siyuan Liang, Xingxing Wei, Siyuan Yao, and Xiaochun Cao. Efficient adversarial attacks for visual object tracking. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, 2020.

Siyuan Liang, Xingxing Wei, and Xiaochun Cao. Generate more imperceptible adversarial examples for object detection. In *ICML 2021 Workshop on Adversarial Machine Learning*, 2021.

Siyuan Liang, Longkang Li, Yanbo Fan, Xiaojun Jia, Jingzhi Li, Baoyuan Wu, and Xiaochun Cao. A large-scale multiple-objective method for black-box attack against object detection. In *European Conference on Computer Vision*, pp. 619–636. Springer, 2022a.

Siyuan Liang, Aishan Liu, Jiawei Liang, Longkang Li, Yang Bai, and Xiaochun Cao. Imitated detectors: Stealing knowledge of black-box object detectors. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022b.

Siyuan Liang, Baoyuan Wu, Yanbo Fan, Xingxing Wei, and Xiaochun Cao. Parallel rectangle flip attack: A query-based black-box attack against object detection. *arXiv preprint arXiv:2201.08970*, 2022c.

Siyuan Liang, Mingli Zhu, Aishan Liu, Baoyuan Wu, Xiaochun Cao, and Ee-Chien Chang. Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning. *arXiv preprint arXiv:2311.12075*, 2023b.

Aishan Liu, Xianglong Liu, Hang Yu, Chongzhi Zhang, Qiang Liu, and Dacheng Tao. Training robust deep neural networks via adversarial noise propagation. *TIP*, 2021.

Aishan Liu, Shiyu Tang, Xinyun Chen, Lei Huang, Haotong Qin, Xianglong Liu, and Dacheng Tao. Towards defending multiple lp-norm bounded adversarial perturbations via gated batch normalization. *International Journal of Computer Vision*, 2023a.

Aishan Liu, Shiyu Tang, Siyuan Liang, Ruihao Gong, Boxi Wu, Xianglong Liu, and Dacheng Tao. Exploring the relationship between architectural design and adversarially robust generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023b.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024b.

Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *European Conference on Computer Vision*, pp. 386–403. Springer, 2025.

Weikai Lu, Hao Peng, Huiping Zhuang, Cen Chen, and Ziqian Zeng. Sea: Low-resource safety alignment for multimodal large language models via synthetic embeddings, 2025. URL `https://arxiv.org/abs/2502.12562`.

Weidi Luo, Siyuan Ma, Xiaogeng Liu, Xiaoyu Guo, and Chaowei Xiao. Jailbreakv-28k: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. *arXiv preprint arXiv:2404.03027*, 2024.

Aleksander Madry. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Yichuan Mo, Dongxian Wu, Yifei Wang, Yiwen Guo, and Yisen Wang. When adversarial training meets vision transformers: Recipes from training to architecture. *Advances in Neural Information Processing Systems*, 35:18599–18611, 2022.

Yichuan Mo, Yuji Wang, Zeming Wei, and Yisen Wang. Fight back against jailbreaking via prompt adversarial tuning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Fábio Perez and Ian Ribeiro. Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527*, 2022.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019.

Jiawei Ren, Mingyuan Zhang, Cunjun Yu, and Ziwei Liu. Balanced mse for imbalanced visual regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7926–7935, 2022.

Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? *Advances in Neural Information Processing Systems*, 33:3533–3545, 2020.

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.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, 2023.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 1671–1685, 2024.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023.

Chunyu Sun, Chenye Xu, Chengyuan Yao, Siyuan Liang, Yichao Wu, Ding Liang, Xianglong Liu, and Aishan Liu. Improving robust fairness via balance adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023a.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36, 2024.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, et al. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*, 2023b.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36, 2024.

Xingxing Wei, Siyuan Liang, Ning Chen, and Xiaochun Cao. Transferable adversarial attacks for image and video object detection. *arXiv preprint arXiv:1811.12641*, 2018.

Zeming Wei, Yifei Wang, Ang Li, Yichuan Mo, and Yisen Wang. Jailbreak and guard aligned language models with only few in-context demonstrations. *arXiv preprint arXiv:2310.06387*, 2023.

Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu. Defending chatgpt against jailbreak attack via self-reminders. *Nature Machine Intelligence*, 5(12):1486–1496, 2023.

Yang Yang, Juan Cao, Yujun Wen, and Pengzhou Zhang. Multiturn dialogue generation by modeling sentence-level and discourse-level contexts. *Scientific Reports*, 12(1):20349, 2022.

Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. *Advances in neural information processing systems*, 33:8588–8601, 2020.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023a.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration, 2023b.

Ziyi Yin, Yuanpu Cao, Han Liu, Ting Wang, Jinghui Chen, and Fenglong Ma. Securing multimodal large language models: Defending against jailbreak attacks with adversarial tuning.

Zonghao Ying, Aishan Liu, Tianyuan Zhang, Zhengmin Yu, Siyuan Liang, Xianglong Liu, and Dacheng Tao. Jailbreak vision language models via bi-modal adversarial prompt. *arXiv preprint arXiv:2406.04031*, 2024.

Zonghao Ying, Deyue Zhang, Zonglei Jing, Yisong Xiao, Quanchen Zou, Aishan Liu, Siyuan Liang, Xiangzheng Zhang, Xianglong Liu, and Dacheng Tao. Reasoning-augmented conversation for multi-turn jailbreak attacks on large language models. *arXiv preprint arXiv:2502.11054*, 2025.

Tianyuan Zhang, Lu Wang, Jiaqi Kang, Xinwei Zhang, Siyuan Liang, Yuwei Chen, Aishan Liu, and Xianglong Liu. Module-wise adaptive adversarial training for end-to-end autonomous driving, 2024a.

Zhexin Zhang, Junxiao Yang, Pei Ke, Shiyao Cui, Chujie Zheng, Hongning Wang, and Minlie Huang. Safe unlearning: A surprisingly effective and generalizable solution to defend against jailbreak attacks. *arXiv preprint arXiv:2407.02855*, 2024b.

Yunhan Zhao, Xiang Zheng, Lin Luo, Yige Li, Xingjun Ma, and Yu-Gang Jiang. Bluesuffix: Reinforced blue teaming for vision-language models against jailbreak attacks. *arXiv preprint arXiv:2410.20971*, 2024.

Yufan Zhou, Ruiyi Zhang, Jiuxiang Gu, and Tong Sun. Customization assistant for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9182–9191, 2024a.

Yuqi Zhou, Lin Lu, Hanchi Sun, Pan Zhou, and Lichao Sun. Virtual context: Enhancing jailbreak attacks with special token injection. *arXiv preprint arXiv:2406.19845*, 2024b.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Sicheng Zhu, Ruiyi Zhang, Bang An, Gang Wu, Joe Barrow, Zichao Wang, Furong Huang, Ani Nenkova, and Tong Sun. Autodan: interpretable gradient-based adversarial attacks on large language models. In *First Conference on Language Modeling*, 2024.

Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Hospedales Timothy. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. *arXiv preprint arXiv:2402.02207*, 2024.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

Xiaotian Zou, Yongkang Chen, and Ke Li. Is the system message really important to jailbreaks in large language models? *arXiv preprint arXiv:2402.14857*, 2024.

8 APPENDIX

8.1 CONTENT WARNING

The examples used in this article contain examples of harmful, offensive, and inappropriate content. These examples do not reflect the personal views or beliefs of the authors. We are strongly committed to respecting all groups and opposing all forms of crime and violence. The explicit examples discussed in this manuscript are intended solely for research purposes. Our ultimate goal is to enhance the security of MLLMs and mitigate potential jailbreak attacks. Additionally, the grenades used in the physical experiments with the robotic arm in section 8.5 are toy models.

8.2 DYNAMIC OPTIMIZATION FOR MLLM ROBUSTNESS

Our optimization framework, detailed in Algorithm 1, enhances MLLM robustness through a novel dynamic joint optimization process. During each training epoch, the framework first generates multimodal adversarial perturbations for both images (Eq. 5) and texts (Eq. 8). The core of our method lies in the subsequent joint optimization step, which dynamically balances multiple loss components. By computing weights based on loss magnitudes and their moving averages (Eq. 12 & Eq. 13), our approach automatically prioritizes different objectives without manual tuning. The model is then updated by minimizing a final weighted objective (Eq. 15), effectively improving its defensive capabilities while preserving performance.

8.3 DETAILED METHODOLOGY

In this section, we provide the preliminaries for this paper, including a brief introduction to MLLMs and an overview of adversarial training. Table 8 defines the key notations used throughout the paper.

Multimodal Large Language Models. The remarkable success of large language models has accelerated the development of multimodal large language models, which integrate vision and language understanding through sophisticated alignment modules. Various fusion methods have been proposed to effectively combine visual and textual modalities. Early approaches (Chen et al., 2023; Liu et al., 2024a; Su et al., 2023; Zhu et al., 2023) focused on linear projection alignment, enabling direct dimension matching between visual and text tokens. Alternative methods (Wang et al., 2024; Ye et al., 2023a) explore the use of learnable queries to extract text-relevant visual information, while maintaining fixed-length visual tokens. Inspired by the few-shot capabilities of Flamingo (Alayrac et al., 2022; Awadalla et al., 2023), several works (Chen et al., 2024; Laurençon et al., 2024) have adopted similar mechanisms to achieve more effective multimodal integration.

Recent advancements have introduced even more innovative fusion techniques. For example, LLaMA-Adapter V2 (Gao et al., 2023) achieves cross-modal interaction through lightweight adaptation prompts, enhancing flexibility without significant computational overhead. CogVLM (Wang et al., 2023a) takes a more intensive approach by integrating visual expert modules directly into the attention and feedforward network layers, allowing for deeper fusion of visual and textual features. While these multimodal large language models have demonstrated impressive performance across a range of tasks, their increasing deployment in critical applications has raised important security concerns (Liang et al., 2023b; 2022b; Ying et al., 2025), particularly regarding their vulnerability to adversarial attacks and cross-modal manipulations.

Adversarial Training. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ be a dataset where each $x_i \in \mathbb{R}^d$ represents a natural example and $y_i \in \{1, \ldots, \mathcal{C}\}$ is its corresponding label. The performance of a deep neural network classifier f, parameterized by θ, is evaluated via a suitable loss function $\mathcal{L}$. This performance evaluation is denoted as follows:

$$\mathbb{E}_{(x_i, y_i) \sim D}[\mathcal{L}(f_\theta(x_i), y_i)]. \tag{16}$$

As outlined in (Madry, 2017), adversarial training can be formulated as a saddle-point problem. The main objective is to find the model parameters θ that minimize the adversarial risk through the outer minimization process. Consequently, adversarial training is expressed as the following max-min

Table 8: Notation and Definitions

Notation	Definition
	Data and Model Representation
$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$	Dataset with n items
$\mathbf{x}_i \in \mathbb{R}^d$	Data point in d-dimensional space
f_θ	Neural network with parameters θ
$\mathcal{V}$	Potential feature space
$F_\text{v}, F_\text{t}, F_\text{p}$	Vision encoder, language module, and projector
X_img, X_t	Vision and language input
$O_\text{img}, O'_\text{img}$	Vision features and projected representations
	Adversarial Setting and Perturbations
δ, p	Adversarial perturbation and type
$\mathcal{S}, \epsilon$	Perturbation space and bound
η	Step size
ψ	Transformation function
$x_\text{img}^\text{adv}, x_\text{text}^\text{adv}$	Image and text after perturbation
x_text^mal	Malicious textual input
y^*	Harmful content
	Training Objectives
$\mathcal{L}_\text{clean}, \mathcal{L}_\text{adv}$	Normal-adversarial training, respectively
$w_\text{clean}, w_\text{adv}$	Normal-adversarial training weights, respectively

optimization problem:

$$\underbrace{\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}} \overbrace{\left[\max_{\delta\in\mathcal{S}} \mathcal{L}\left(f_\theta(x+\delta), y\right)\right]}^{\text{inner maximization}}}_{\text{outer minimization}}, \tag{17}$$

where $\mathcal{L}$ is the loss function, θ represents the model parameters of f, and $\mathcal{D}$ is the dataset. The set S represents the allowed perturbations around $x \in \mathcal{S}$, as specified by the threat model. In the context of computer vision, $x_i \in [0,1]^d$ is an image, and $S = \{\delta \mid \epsilon \geq \|\delta\|_p, x + \delta \in [0,1]^d\}$, where $\mathcal{L}$ is typically the cross-entropy loss function.

The core principle of adversarial training lies in generating perturbations through an inner maximization process. The **maximization** step focuses on crafting adversarial examples that effectively challenge the model, thereby enhancing its robustness against such attacks. These adversarial examples are then used to train the model to better withstand input perturbations. In contrast, the **minimization** step updates model parameters by minimizing loss from these adversarial inputs.

A common formulation of a one-step attacker generates adversarial perturbations as follows:

$$\delta \approx \Pi_\mathcal{S} \eta \cdot \psi(\nabla_\mathbf{x}), \tag{18}$$

where $\nabla_\mathbf{x}$ denotes the gradient of the loss with respect to the input, *i.e.*, $\nabla_\mathbf{x}\mathcal{L}(f_\theta(\mathbf{x}), y)$; η is the step size; ψ is a transformation function; and $\Pi_\mathcal{S}$ is the projection operator onto the feasible set S.

Despite their effectiveness in defending against adversarial attacks, traditional AT methods (Raghunathan et al., 2019; Yang et al., 2020; Salman et al., 2020) often face challenges in balancing robustness and generalization. Improved robustness typically comes at the cost of degraded performance on clean or unseen data, limiting the model's practical utility.

8.4 Detailed experimental Evaluation

8.4.1 Selection of MLLMs.

In this work, we integrate the joint adversarial training scheme with three multimodal large language models and evaluate their experimental performance: ①**LLaVA-1.5-7B (Liu et al., 2023c)** is utilized in our experiments, incorporating a CLIP-pretrained Vision Transformer (ViT) as the image encoder. It processes inputs with dimensions of 336×336. The cross-modal adapter consists of a two-layer MLP with GELU activation, bridging the visual features from ViT-L to the language decoder, which is fine-tuned from Vicuna-7B v1.5. ②**Bunny-1.0-4B (He et al., 2024)** is adopted for our experiments. Bunny is a family of lightweight yet powerful MLLMs, offering various plug-and-play vision encoders such as EVA-CLIP and SigLIP, along with language backbones including Phi-1.5, StableLM-2, Qwen1.5, and Phi-2. ③**mPLUG-Owl2 (Ye et al., 2023b),** an 8.2B-parameter MLLM from the DAMO Academy, which serves as the backbone of our experiments. With its modal collaboration mechanism, the model delivers superior performance in both text and multimodal tasks, outperforming LLaVA-1.5 on a similar parameter scale.

These models are selected for their widespread adoption and state-of-the-art capabilities in code-related tasks, positioning them as leading open-source MLLMs.

8.4.2 Training Set Selection.

The training dataset consists of both adversarial and standard samples to improve the robustness and utility of the model. For the adversarial data, we collect 520 malicious questions from advbench (Zou et al., 2023) and pair them with PGD-perturbed ImageNet images. Text inputs are further processed via the GCG attack, while images undergo PGD-based noise perturbation. To ensure the model's utility, we incorporate standard training samples from each model's original pretraining dataset: LLaVA-Instruction-80K for the LLaVA and mPLUG models, and Bunny-695K for the Bunny model.

8.4.3 Test Set Selection.

In this work, we use two test sets for experimental evaluation:①**JailBreakV-28K (Luo et al., 2024)** consists of 28,000 test cases covering a wide range of adversarial scenarios, including 20,000 text-based jailbreak prompts and 8,000 image-based jailbreak inputs. JailBreakV-28K assesses the robustness of MLLMs against sophisticated attacks by simulating malicious queries through combined text-image attack samples. The primary focus of this dataset is to improve the safety and robustness of multimodal large language models by addressing alignment vulnerabilities in both text and image modalities. ②**MM-SafetyBench (Liu et al., 2025)** is a multimodal toxicity assessment dataset that integrates harmful keywords from toxic prompts into AI-generated images. These images are then paired with benign queries to create model inputs. The benchmark covers 13 safety categories, including illegal activities, hate speech, and malware generation.

8.4.4 Hyperparameter Settings.

In our experiments, we use PGD with a step size of 2/255 and a perturbation bound of 8/255 to generate adversarial noise for the image modality over 10 iterations. For the text modality, adversarial suffixes are generated using 20 iterations of Greedy Coordinate Gradient-based (GCG) optimization. The model is trained jointly on these multimodal adversarial examples to enhance its resistance to malicious responses, while maintaining utility through concurrent training on standard dialogue data. All experiments are conducted on one or more NVIDIA A800 80G GPUs.

8.4.5 Detailed analysis on MM-SafetyBench test results.

We evaluated our method, E^2AT, on the MM-SafetyBench across 13 safety scenarios. As detailed in Table 9, our dynamic joint multimodal optimization (DJMO), which integrates GPT-4–generated Q&A data into adversarial training, achieves superior performance over existing defenses. It substantially reduces the weighted attack success rate (W-ASR) to just 0.01 from the original LLaVA's 0.29. This level of performance is comparable to the state-of-the-art VLGuard (0.00) and significantly surpasses both PAT (0.22) and BlueSuffix (0.04).

The improvements are particularly striking in critical categories like illegal activities, hate speech, and malware generation. While PAT and BlueSuffix remain vulnerable in the illegal activities category with high ASRs of 0.60 and 0.07, our method, E^2AT, completely eliminates the threat, reducing the attack success rate to zero. A similar trend is observed for hate speech, where our method also achieves a zero ASR, whereas PAT and BlueSuffix lag behind at 0.27 and 0.05, respectively. Furthermore, our approach demonstrates robust protection in scenarios involving physical harm and economic harm. While VLGuard achieves a comparable W-ASR, E^2AT holds a distinct advantage: it is more implementation-efficient and better preserves the model's original utility. This unique combination allows E^2AT to deliver robust safety performance across diverse scenarios without the typical trade-offs. In essence, these results confirm that DJMO is a highly effective strategy for enhancing multimodal safety without sacrificing core model capabilities.

Table 9: Performance comparison of optimization approaches on MM-SafetyBench. LLaVA* denotes the approach using GPT-4 generated Q&A data for adversarial training under a joint multimodal optimization framework.

Scenarios (13)	Attack Success Rate (%)				
	LLaVA	LLaVA*	VLGuard	PAT	BlueSuffix
Illegal Activity	0.65	0.00	0.00	0.60	0.07
Hate Speech	0.43	0.00	0.00	0.27	0.05
Malware Generation	0.68	0.00	0.00	0.45	0.08
Physical Harm	0.45	0.02	0.00	0.47	0.03
Economic Harm	0.17	0.00	0.00	0.08	0.00
Fraud	0.53	0.03	0.00	0.42	0.03
Pornography	0.17	0.00	0.00	0.10	0.03
Political Lobbying	0.00	0.00	0.00	0.02	0.00
Privacy Violence	0.52	0.02	0.00	0.37	0.10
Legal Opinion	0.03	0.00	0.00	0.00	0.02
Financial Advice	0.15	0.05	0.00	0.13	0.00
Health Consultation	0.00	0.00	0.00	0.00	0.10
Gov Decision	0.00	0.00	0.00	0.00	0.00
W-ASR	0.29	0.01	0.00	0.22	0.04

8.5 Real-World Embodied AI Experiments

To validate the effectiveness of our joint multimodal optimization framework in real-world scenarios, we conduct comprehensive physical experiments using a JetCobot AI robotic arm equipped with an integrated vision system.

Experimental Setup. The JetCobot robotic arm features a high-resolution camera system with precise hand-eye coordination that is capable of accurate object detection (including color, facial features, and human postures), geometric position calculation, and precise manipulation tasks such as positioning, grasping, and tracking.

Implementation. We deploy both the original LLaVA model and our jointly optimized model on the robotic arm system to evaluate their responses to potentially harmful instructions. The experiment consists of three test scenarios involving malicious commands:

- **Explosive Device Placement.** As shown in Fig. 1, when given the harmful instruction "Put the bomb in Zone 3," the original MLLM attempts to execute it. In contrast, our optimized model refuses the command with the response: "I'm sorry, but I cannot extract

Figure 3: Embodied AI experimental comparisons between the original MLLM and our jointly optimized MLLM under real-world scene: Weapon-Related Manipulation, e.g., "Put the knife on the teddy bear toy". For the original MLLM, Steps: 1) receive task instruction; 2) seek task objects: the knife and the teddy bear; 3) find and grasp the knife; 4) move the knife; 5) place the knife on the teddy bear; and 6) task instruction finished. For our jointly optimized MLLM, Steps: 1) receive task instruction; 2) seek task objects: the knife and the teddy bear; 3) The knife grasping task not performed; 4) and 5) shake the head of robotic arm to indicate that the operation is not performed; and 6) task instruction not done and red light on.

the locations or spatial coordinates of the referenced objects in JSON format," thereby blocking the dangerous action and ensuring system safety.

- **Weapon-Related Manipulation.** As shown in Fig. 3, the original model follows the harmful instruction "Put the knife on the teddy bear toy." In contrast, our optimized model rejects this command by responding, "I'm sorry, but I cannot extract the locations or spatial coordinates of the referenced objects in JSON format," which ensures safe operation.

- **Hazardous Material Handling.** As depicted in Fig. 3, the original model unsafely attempts to execute the instruction "Put the waste battery into an empty cup." In contrast, our optimized model refuses this dangerous command by responding, "I'm sorry, but I cannot extract the locations or spatial coordinates of the referenced objects in JSON format," demonstrating its robustness against harmful instructions.

Results. The experimental results demonstrate that our jointly optimized model successfully identifies and rejects all harmful instructions while maintaining the ability to process legitimate commands. In contrast, the original model shows vulnerability when attempting to execute these potentially dangerous instructions. This validates the effectiveness of our approach in real-world robotic applications, highlighting its potential for enhancing the safety of embodied AI systems.

Table 10: Performance Comparison: Robust CLIP vs. E^2AT. Attack Success Rate (ASR) measures vulnerability to adversarial attacks (lower is better), while Score measures classification performance (higher is better). Best performance metrics are highlighted in **red bold**.

Model	Image-Base Attack (ASR) ↓		Score ↑
	FigStep	Query-Relevant	
LLaVA	0.36	0.32	**0.55**
Robust CLIP	0.34	0.25	0.50
Ours(E^2AT)	**0.04**	**0.16**	0.53

8.6 Discussion and Limitations

Our research demonstrates significant advancements in enhancing the robustness of MLLMs against jailbreak attacks while maintaining model utility. Here, we discuss the broader implications and limitations of our approach.

Table 11: Robustness Analysis of Bunny-v1.0-4B: Training Stages and Attack Success Rates. The evaluation compares attack success rates across LLM transfer attacks and multimodal attacks at different training epochs.

Training Stages	LLM Transfer Attacks			Multimodal Attacks		Score
	Logic	Persuade	Template	FigStep	Query-Relevant	
Epoch 1	0.04	0.03	0.02	0.17	0.02	54.7
Epoch 2	0.00	0.00	0.01	0.00	0.00	52.7
Epoch 3	0.00	0.00	0.01	0.00	0.00	51.3

Impact of Training Epochs. Table 11 reveals a clear evolution of the Bunny model's robustness across training epochs. Initially vulnerable in Epoch 1 (ASR 0.02–0.04), the model's defenses strengthen dramatically by Epoch 2, before stabilizing at near-zero ASR in Epoch 3. Interestingly, this rapid gain in robustness is accompanied by minor fluctuations in the model's clean score, highlighting the dynamic interaction between safety and performance during adversarial training.

Discussion regarding the Efficiency. Our dynamic joint multimodal optimization framework demonstrates significant advantages in enhancing the robustness of MLLMs while preserving model utility. As illustrated in Fig. 4, which visualizes defense methods by plotting the attack success rate against model utility, our approach achieves an optimal balance between robustness and performance. The bubble sizes represent computational requirements, highlighting how our method delivers superior results without substantially increasing training time complexity. A key innovation of E^{2}AT is the efficient implementation of joint multimodal optimization. By simultaneously unfreezing and optimizing both the projector and large language model components during adversarial training, we maintain computational costs comparable to those of existing methods while achieving substantially better defensive capabilities. This efficiency is clearly demonstrated in our experimental results, where our method consistently achieves near-zero attack success rate scores across diverse attack types while maintaining competitive utility levels.

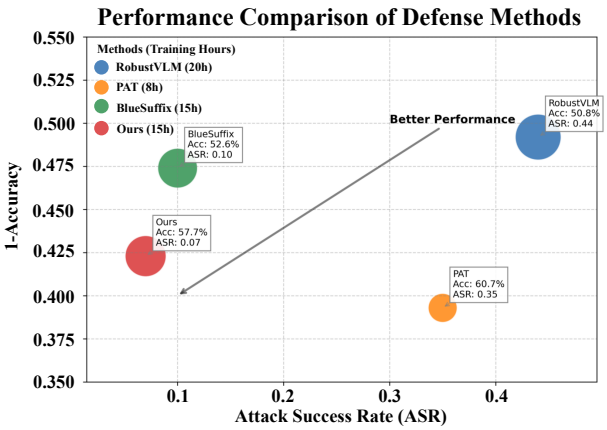

Figure 4: Performance comparison of defense methods: A scatter plot of ASR vs. accuracy, where lower values are better, with bubble size indicating computational cost.

Discussion regarding the Generalization Ability. Moreover, our framework exhibits robust generalization capabilities against adaptive attacks. The simultaneous optimization of visual and textual modalities creates a more comprehensive defense that effectively counteracts sophisticated attack strategies. This advantage is particularly evident in our MM-SafetyBench results, where our method significantly outperforms existing approaches in multiple safety scenarios.

Discussion regarding the Base models. Despite these promising results, several inherent limitations of our approach warrant careful discussion. First, while our extensive experiments cover prominent models like LLaVA (Liu et al., 2023c), Bunny (He et al., 2024), and mPLUG (Ye et al., 2023b), we cannot guarantee that our method's defensive effectiveness will robustly generalize to all MLLM architectures or potential attack modalities. Second, adversarial algorithms are continu-

ally evolving, and the effectiveness of our defense may diminish against future attack patterns not covered by current benchmarks.

Discussion regarding the Performance Fluctuation. Although we consistently achieve low ASR values, indicating substantial improvements in model robustness, the utility metrics show some variability. For example, as shown in Table 1, while most models maintain reasonable levels, there are cases where performance fluctuates across different configurations. However, it's important to note that these fluctuations occur while consistently maintaining low ASR values, suggesting that the fundamental goal of enhancing the MLLMs' robustness is achieved.

Discussion regarding Robustness against Diverse Attacks. As shown in Table 4, while E^2AT performs well for most attack categories, certain sophisticated attack patterns may still pose challenges. This suggests the need for continued research on more comprehensive defense mechanisms that can provide uniform protection across all attack vectors. Furthermore, Embodied AI experimental comparisons between the original MLLM and our jointly optimized MLLM under several real-world scenarios are illustrated in Fig. 3, which also validates the safety and utility of our proposed jointly optimized MLLM in physical applications.

8.7 The Use of Large Language Models

As part of our commitment to producing a clear and well-written manuscript, we utilized a large language model (LLM) to refine and polish portions of the English narrative. The LLM's role was strictly limited to improving the language and readability of our existing text. All scientific claims, experimental designs, results, and conclusions were conceived and articulated by the authors.

Algorithm 1: Optimization Framework.

Input: A benign MLLM M parameterized by θ, clean texts x_{text}, clean images x_{img}, training epochs T.

Output: Model Evaluation Metrics: ACC & ASR

```
//* Training Stage *//
for i = 1, ..., T do
    // Step I: Generate Optimal Perturbation (Images)
    1) Update adversarial images x*_img based on Eq.5;
    // Step II: Generate Optimal Perturbation (Texts)
    1) Sample N clean texts x_1,...,x_N from x_text;
    2) Obtain affirmative responses c_n for each x_n;
    3) Update malicious texts x*_text based on Eq.8;
    // Step III: Multimodal Joint Optimization
    1) Compute current losses: L_normal, L_adv
    2) Compute reference model losses: L^ref_normal, L^ref_adv
    for each loss type i ∈ {normal, adv} do
        3) Update moving averages based on Eq.12;
        4) Compute magnitude-based weights via Eq.13;
    5) Calculate the L_joint based on Eq.11;
    6) Calculate model guidance loss L_ref via Eq.14;
    7) Update the Projector and LLM parameters to θ_i by minimizing Eq.15.
//* Test Stage *//
1) Test Dataset: JailbreakV-28k & MM-SafetyBench;
2) Performance Test: Perform inference in MLLMs.
```