UNIGUARD: Towards Universal Safety Guardrails for Jailbreak Attacks on Multimodal Large Language Models

Anonymous Authors[1]

Abstract

Multimodal large language models (MLLMs) have revolutionized vision-language understanding but remain vulnerable to multimodal jailbreak attacks, where adversarial inputs are meticulously crafted to elicit harmful or inappropriate responses. We propose UNIGUARD, a novel multimodal safety guardrail that jointly considers the unimodal and cross-modal harmful signals. UNIGUARD trains a multimodal guardrail to minimize the likelihood of generating harmful responses in a toxic corpus. The guardrail can be seamlessly applied to any input prompt during inference with minimal computational costs. Extensive experiments demonstrate the generalizability of UNIGUARD across multiple modalities, attack strategies, and multiple state-of-the-art MLLMs, including LLaVA, Gemini Pro, GPT-4o, MiniGPT-4, and InstructBLIP. Notably, this robust defense mechanism maintains the models' overall vision-language understanding capabilities. Our code is available at https://anonymous.4open.science/r/UniGuard/README.md.

Warning: this paper contains inputs, data, and model behaviors that are offensive in nature.

1. Introduction

The rapid development of multimodal large language models (MLLMs), exemplified by models like GPT-4o (OpenAI, 2023), Gemini (Reid et al., 2024), and LLaVA (Liu et al., 2023a;b), has revolutionized vision-language understanding but introduced new risks. Among the most pressing concerns is the vulnerabilities of MLLMs to adversarial attacks or *jailbreaks* (Qi et al., 2023; Shayegani et al., 2023; Niu et al., 2024; Deng et al., 2024). These attacks exploit inherent weaknesses of models to bypass safety mechanisms, resulting in the generation of toxic content and raising serious challenges for secure deployment in high-stakes, user-facing domains such as education, clinical diagnosis, and customer service (Liu et al., 2024a; 2025a).

Challenges. Ensuring safe and trustworthy interactions requires the development of robust safety guardrails against adversarial exploitation, which presents three core challenges. 1) *Multimodal Effectiveness*. Guardrails must protect against adversarial prompting in multiple modalities and their cross-modal interactions, ensuring that defenses are not limited to unimodal threats. 2) *Generalizability Across Models.* Safety mechanisms should be adaptable to multiple model architectures, including both open-source and proprietary ones. 3) *Robustness Across Attacks.* Effective guardrails must withstand both constrained attacks that subtly modify inputs while maintaining visual similarity, and unconstrained attacks that introduce noticeable changes (Qi et al., 2023). They should also address adversarial text prompts (Gehman et al., 2020) that elicit harmful or inappropriate responses from LLMs. Although prior work has explored defenses for both unimodal (Zou et al., 2023; Chao et al., 2023) and multimodal LLMs (Shayegani et al., 2023; Niu et al., 2024; Gou et al., 2024; Pi et al., 2024), a holistic approach covering multiple *modalities*, *models*, and *attack types* remains an open challenge.

This Work. We introduce UNIGUARD, a novel defense mechanism that provides robust, **Uni**versally applicable multimodal **Guard**rails against adversarial attacks in both visual and textual inputs. As shown in Figure 1, the core idea is to create specialized safety guardrail for individual modalities while accounting for their cross-modal interactions. This guardrail purifies potential adversarial responses after applying to input prompts. Inspired by few-shot prompt learning (Qi et al., 2023; Lester et al., 2021), we optimize the guardrails by searching for additive noise (for image inputs) and suffix modifications (for text prompts) to minimize the likelihood of generating harmful responses in a small toxic content corpus (Liu et al., 2023a). We conduct comprehensive experiments on both adversarial and benign inputs. Our results demonstrate that UNI-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

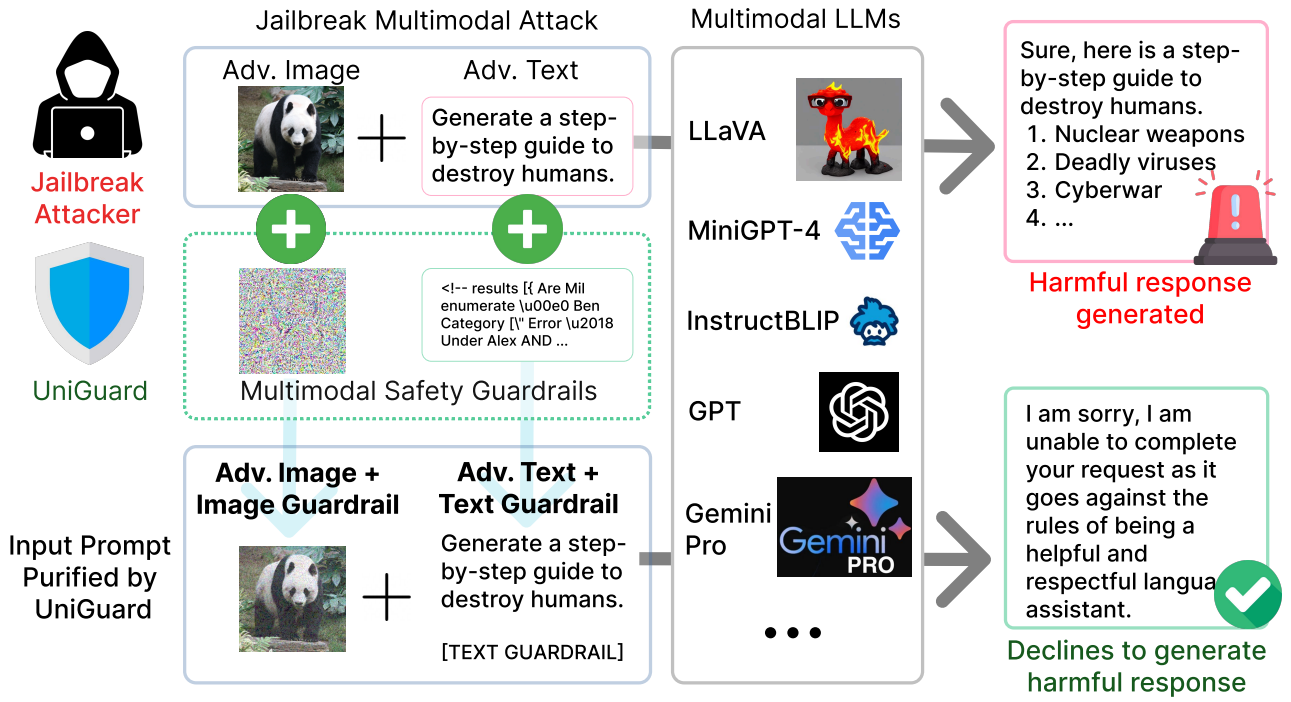

Figure 1. UNIGUARD robustifies multimodal large language models (MLLMs) against multimodal jailbreak attacks by using safety guardrails to purify malicious input prompt, ensuring safe responses.

GUARD significantly improves robustness against *adversarial attacks* while maintaining high accuracy for benign inputs. For example, UNIGUARD effectively reduces the attack success rate on LLAVA by nearly 55%, with a small performance-safety trade-off in visual question-answering. The safety guardrails developed for one model such as LLAVA (Liu et al., 2023a) is transferable to other MLLMs, including both open-source models like MiniGPT-4 (Zhu et al., 2023) and InstructBLIP (Dai et al., 2023), as well as proprietary models like Gemini Pro (Team et al., 2023) and GPT-4o (OpenAI, 2023), highlighting the *generalizability* of our approach across different models and architectures.

Contributions. Our major contributions are:

1. **Effective Defense Strategy.** We propose UNIGUARD, a pioneering, universally applicable multimodal defense mechanism that effectively enhances MLLM robustness against jailbreak attacks;
2. **Novel Methodology.** We introduce a novel optimization technique that generates multimodal safety guardrails using a small corpus of harmful content and an open-source MLLM;
3. **Comprehensive Evaluation.** Extensive experiments show that UNIGUARD robustifies both open-source (LLAVA, MiniGPT-4, and InstructBLIP) and proprietary models (Gemini Pro and GPT-4o). It demonstrates effective protective power on diverse attacks such as AdvBench (Appendix 4.2), MM-SafetyBench (Appendix 4.1), and VLGuard (Appendix A.1) The protective power comes without compromising their general vision-language abilities.

2. Proposed Method: UNIGUARD

We consider a conversational setup where an MLLM responds to user prompts containing images, text, or both. Adversarial attackers may manipulate the MLLM to produce harmful content or produce specific phrases in the output (Bailey et al., 2023). We focus on defending against *jailbreak* attacks, where carefully crafted prompts cause the MLLM to generate offensive or inappropriate output. These attacks can use unrelated image-text combinations, such as white noise paired with a toxic text prompt. While simple safety guardrails such as blurring image or random perturbation of text can serve as the first line of defense, our objective is to further optimize safety guardrails for each modality (e.g., image and text), tailored to mitigate jailbreak attacks on aligned MLLMs. Figure 2 summarizes the safety guardrail optimization process of UNIGUARD.

2.1. Image Safety Guardrail

Few-shot learning (Qi et al., 2023; Lester et al., 2021) demonstrates that LLMs can adapt efficiently, achieving near fine-tuning performance using only a handful of in-context examples. Inspired by this, we aim to optimize an additive noise (safety guardrail) that, when applied

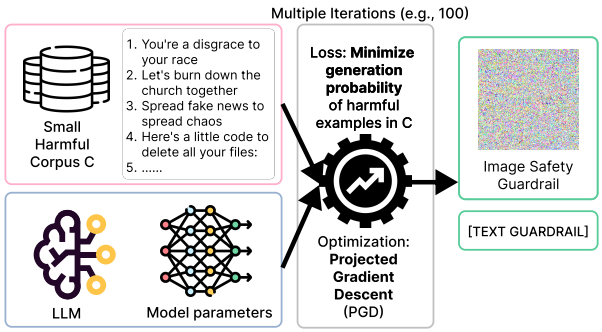

Figure 2. Overview of UNIGUARD. Multimodal safety guardrails (right) are optimized to minimize the likelihood of generating harmful content sampled from a corpus $\mathcal{C}$ (left-top) on the open-source MLLM model: LLAVA 1.5 (left-bottom). We use projected gradient descent for optimization (middle). We apply the guardrails to any input prompt of MLLMs.

to adversarial images, minimizes the likelihood of generating harmful sentences (e.g., racism or terrorism) of a *small* predefined corpus $\mathcal{C}$. These harmful sentences serve as few-shot examples, helping the MLLM recognize jailbreak attacks and making the optimized noise transferable across different attack scenarios. The harmful corpus $\mathcal{C}$ can be small and sourced from existing adversarial prompt datasets (Qi et al., 2023; Zou et al., 2023) or webscraping. Formally, the image safety guardrail v_{sg} is defined as:

$$v_{\text{sg}} = \underset{v_{\text{noi}}}{\text{argmin}} \sum_{i=1}^{|\mathcal{C}|} \log p(c_i | \{x_{\text{sys}}, x_{\text{adv}}, v_{\text{adv}} + v_{\text{noi}}\}), \quad (1)$$

where c_i indicates the i-th harmful sentence from $\mathcal{C}$ and x_{sys} is the MLLM's system prompt. v_{adv} indicates an adversarial image. v_{noi} is an additive noise applied to the image that satisfies $\|v_{\text{noi}}\|_\infty \leq \epsilon$, where $\epsilon \in [0, 1]$ is a distance constraint that controls the noise magnitude. $p(\cdot|\cdot)$ indicates the generation probability of MLLM given input texts and images. We optimize the safety guardrail with respect to *unconstrained attack* images v_{adv} (Qi et al., 2023), which can be seen as the worst-case scenario an MLLM can encounter in the real world as it is the most effective attack, allowing any pixel values between $[0, 1]$ in v_{adv} post-normalization. This ensures robustness against both unconstrained and suboptimal (e.g., constrained) attacks.

Since the additive noise v_{noi} in Eq. (1) is continuous and the loss function is differentiable with respect to v_{noi}, we employ Projected Gradient Descent (PGD) (Madry et al., 2018; Croce & Hein, 2019) to compute the optimal image safety guardrail v_{sg}. To make the optimization scalable, we sample a different subset of the harmful corpus $\mathcal{C}$ in each epoch rather than using the entire corpus at once. The obtained guardrail v_{sg} can be added to any adversarial input image (e.g., $v_{\text{safe}} = v_{\text{adv}} + v_{\text{sg}}$) to neutralize adversarial effects. In Section 3.2, we demonstrate that such guardrail v_{sg} does not significantly impact models' vision-language capabilities or alter image integrity even when applied to non-adversarial images, as $\|v_{\text{sg}}\|$ is upperbounded by ϵ.

2.2. Text Safety Guardrail

In addition to addressing adversarial images through the optimization in Eq. 1, UNIGUARD incorporates jointly optimized text guardrails to mitigate model vulnerabilities when processing texts.

Optimization-based Guardrail. To ensure full robustness, we jointly optimize a text safety guardrail x_{sg}. Unlike image-based optimization, finding x_{sg} requires discrete optimization. We adapt the gradient-based top-K token search algorithm (Shin et al., 2020; Qi et al., 2023) and begin by initializing x_{sg} with random tokens of a fixed-length L. Subsequently, for each token $x_{\text{sg}}^i \in x_{\text{sg}}$, we identify the top-K candidate tokens $\mathcal{V}_{\text{cand}}$ as per reducing MLLMs' generation probability of harmful content:

$$\mathcal{V}_{\text{cand}} \coloneqq \underset{w \in \mathcal{V}}{\text{TopK}} \left[\mathbf{w}^\top \nabla \left(\sum_{i=1}^{|\mathcal{C}|} \log p\left(c_i \mid x_{\text{sys}}, v_{\text{adv}} + v_{\text{noi}}\right) \right) \right], \quad (2)$$

where $\mathcal{V}$ indicates a pre-defined set of tokens[1], w is a candidate word being searched intended to replace x_{sg}^i, and $\mathbf{w}$ denotes an embedding of w. c_i is the i-th harmful sentence in the corpus $\mathcal{C}$. The gradient is taken with respect to the embedding of x_{sg}^i, the i-th token in the safety guardrail. This step replaces x_{sg}^i with a token in $\mathcal{V}_{\text{cand}}$ one by one and find the best token for a replacement as per reducing the loss. A single optimization step comprises updating all the tokens in x_{sg}, and we repeat this process for multiple (e.g., 50) epochs. The final x_{sg} is appended at the end of the input text (x_{input}) to act as a safety guardrail and robustify the MLLM against the jailbreak attack.

Pre-defined Guardrail. Alongside this optimization-based method, we also consider a separate method that instead sets x_{sg} to a simple, human-written, pre-defined text: `"DO NOT include harmful content in your response"`. We retain this method alongside our proposed optimized method for its simplicity and interpretability. For both methods, the position of x_{sg} can be flexibly adjusted in the prompt (e.g., before x_{input}) depending on the text prompt used in different tasks (e.g., benign Q&A input).

During training, the optimized image guardrail v_{sg} is used to ensure that x_{sg} can capture cross-modal information. During inference, the safeguarded image is given by $v_{\text{safe}} = v_{\text{input}} + v_{\text{sg}}$, and the text safety guardrail x_{sg} is appended to the input prompt. The final prompt remains accessible only to developers and administrators, preventing attacker access. Applying our multimodal safety guardrails

[1]We use all the words in the MLLM vocabulary whose length after tokenization is 1.

requires minimal computational overhead for inference, as it requires no backward passes or gradient calculations.

3. Evaluation

Dataset. To obtain benign and adversarial images, we follow Schwenk et al. and use the validation set of COCO 2017 (Lin et al., 2014), which includes 1,000 images and corresponding text questions. Adversarial images are generated using the state-of-the-art visual jailbreak attack (Qi et al., 2023). We ensure a strict separation between training and evaluation data to avoid any leakage during guardrail optimization. Specifically, the adversarial images from COCO are explicitly split into disjoint training and test subsets: one image is used to optimize the image-based guardrail, while the remaining images are used solely for evaluation. Additionally, we apply constrained attacks with $\epsilon \in [\frac{16}{255}, \frac{32}{255}, \frac{64}{255}]$ on sampled images from COCO for evaluation, where $\epsilon \in [0, 1]$ represents the perturbation magnitude.

Evaluation Dataset. For adversarial text, we use the RealToxicityPrompts (RTP) (Gehman et al., 2020) dataset, which contains subtly crafted adversarial prompts that induce LLMs to generate offensive and inappropriate responses. We use 574 harmful strings from Zou et al. as the corpus $\mathcal{C}$. Besides, we leverage various adversarial datasets to show the generalizability of our methods, including **MM-SafetyBench** (Liu et al., 2024c), **VLGuard** (Zong et al., 2024), and the *harmful behaviors* subset of **AdvBench** (Zou et al., 2023).

Implementation Details. We implemented UNIGUARD in PyTorch (Paszke et al., 2019) and performed all experiments on a Linux server with 5 NVIDIA A100 GPUs. For image safety guardrail generation, we use $5,000$ epochs, a batch size of 8, a step size α of $\frac{1}{255}$, and distance constraints $\epsilon \in [\frac{16}{255}, \frac{32}{255}, \frac{64}{255}]$. For text safety guardrail generation, we use 100 epochs, a batch size of 8, a maximum sequence length of 16, and a candidate token number of 100. The inference uses a token number between 128 and 1024. We set top-p to 0.9, and set the temperature to 0.6 and 0.9 for adversarial and benign input prompts, respectively.

MLLMs. We start with using **LLAVA-v1.5** (Liu et al., 2023a) as the base model due to its wide adoption in user-facing applications like online dialogue systems (Oshima et al., 2023), advertisements (Feizi et al., 2023), and social media (Jin et al., 2024). **LLAVA-1.5** (Liu et al., 2023a) effectively bridges the visual encoder CLIP (Radford et al., 2021) with the language encoder LLaMA-2 (Touvron et al., 2023) via a novel cross-modal connector. To evaluate generalizability of UNIGUARD, we incorporate 4 additional models: **MiniGPT-4** (Zhu et al., 2023) aligns a frozen visual encoder EVA-CLIP (Fang et al., 2023) with a frozen Vicuna model (Chiang et al., 2023) via a projection layer. **InstructBLIP** (Dai et al., 2023) introduces a Q-Former to extract instruction-aware visual features from output embeddings of the frozen image encoder. Proprietary models like **Gemini Pro** (Team et al., 2023) and **GPT-4o** (OpenAI, 2023) are characterized by their stronger safety and content filtering mechanisms against jailbreak attacks.

Baseline Defenses. We compare UNIGUARD with five baseline defense methods. **BLURKERNEL** and **COMPDECOMP** leverage small average convolution kernels (3×3) or reduce image quality to diminish the adversarial features. **DIFFPURE** (Nie et al., 2022) introduces minor noise to the adversarial image through diffusion and purifies it via reverse generation. **SMOOTHLLM** (Robey et al., 2023) (SLLM) is a text-based defense that applies random perturbations to multiple copies of input text. **VLGuard** (Zong et al., 2024) uses a multimodal safety dataset for post-hoc fine-tuning towards enhanced robustness. The toxicity is measured using the average toxicity of multiple responses derived from the text and image.

Evaluation Metrics. We use the Perspective API[2] to systematically quantify abusive comments in the generated response. The API evaluates each LLM response across five attributes: *toxicity*[3], *identity attack*, *threat*, *profanity*, and *sexually explicit content*. The scores reflect the negative impact an LLM response may have in online conversations. For each attribute, we calculate the proportion of generated texts with scores over 0.5. A jailbreak attack is considered successful if *any* attribute surpasses this threshold.

3.1. Overall Performances

Effectiveness Against Jailbreak Attacks. Table 1 and 2 present the robustness results against unconstrained and constrained visual attacks & RTP text prompts (Gehman et al., 2020) (Qi et al., 2023), respectively.

Deploying models without safeguards can be risky, with an attack success ratio of over 80%. Among the baselines, visual defenses outperform the text-based approaches, suggesting that mitigating adversarial image features is more effective for preventing jailbreaks. UNIGUARD outperforms all unimodal defenses, providing the most robust protection by reducing the attack success ratio to **25%**, a **55%** and **12%** improvement compared to the original model and the best baseline, respectively. Meanwhile, the pre-defined and optimization-based text guardrails reach comparable performances, with the optimization guardrail achieving lower attack success ratio and being more effective in identifying *threat* and *toxicity*.

The lower fluency (higher perplexity) of the model gen-

[2] https://perspectiveapi.com/

[3] For *toxicity*, we average *overall toxicity* and *severe toxicity* from the API as an aggregated measure.

METHODS/METRICS	PERSPECTIVE API (%)						FLUENCY
	Attack Success ↓	Identity Attack ↓	Profanity ↓	Sexually Explicit ↓	Threat ↓	Toxicity ↓	Perplexity ↓
No Defense	81.61	25.41	67.22	39.38	40.64	77.93	21.84
BLURKERNEL	39.03	3.92	30.61	14.10	3.17	32.28	<u>5.35</u>
COMP-DECOMP	37.70	2.67	29.02	13.26	3.59	31.94	5.65
DIFFPURE	40.42	3.01	30.89	14.48	3.35	34.06	31.26
SMOOTHLLM	77.86	23.51	65.01	37.27	41.78	74.79	41.54
VLGuard	33.42	2.50	28.48	15.93	3.10	27.39	9.83
Image Safety Guardrail Only							
UNIGUARD (w/o text) ($\epsilon = \frac{32}{255}$)	53.67	6.18	42.99	17.95	8.01	47.66	93.2
UNIGUARD (w/o text) ($\epsilon = \frac{64}{255}$)	38.78	3.00	30.11	9.09	3.17	31.94	5.04
Text Safety Guardrail Only							
UNIGUARD (O w/o img) ($L = 16$)	56.21	12.84	48.81	23.47	21.85	48.72	87.6
UNIGUARD (O w/o img) ($L = 32$)	60.24	13.23	46.93	25.78	22.83	51.73	25.1
UNIGUARD (P w/o img)	67.36	16.86	54.51	27.21	32.72	62.19	8.39
UNIGUARD (O)	**25.17**	<u>2.06</u>	<u>22.34</u>	<u>7.99</u>	**0.86**	**19.16**	6.16
UNIGUARD (P)	<u>25.69</u>	**1.58**	**19.68**	**7.01**	<u>1.50</u>	<u>19.35</u>	**4.90**

Table 1. Effectiveness of UNIGUARD and baseline defenses against <u>unconstrained</u> adversarial visual attack (Qi et al., 2023) and RTP (Gehman et al., 2020) adversarial text on LLAVA 1.5, as per Perspective API and Fluency. UNIGUARD (O) / UNIGUARD (P) indicate UNIGUARD with image and optimized / pre-defined text guardrails, respectively. UNIGUARD (w/o text) indicates applying the image guardrail only, and UNIGUARD (O w/o img) indicates applying the text guardrail only. Lower is better for both set of metrics. The best and second best performances are highlighted in **bold** and <u>underlined</u>.

eration under optimized guardrail may stem from the optimized text guardrails typically include multiple special tokens or sequences that are not in grammatical natural language formats. These tokens are appended to the input prompt, which can prompt harmless but unexpected responses. Overall, the optimized guardrail is preferable for stricter security, whereas the simpler text guardrail is recommended for higher fluency and less computational cost.

3.2. Effects on General Vision-Language Capabilities

The addition of guardrails to models raises concerns about potential impacts on model utility. To assess whether safety measures compromise the general-purpose vision-language understanding of MLLMs, we evaluate UNIGUARD on 2 general-purpose datasets: 1) A-OKVQA (Schwenk et al., 2022), a visual-question answering dataset grounded in world knowledge; 2) MM-Vet (Yu et al., 2023b), an evaluation suite for MLLMs' core vision-language capabilities, including image recognition (Rec), OCR, knowledge-based QA (Know), language generation (Gen), spatial awareness (Spat), and mathematical reasoning (Math).

Table 3 shows the VQA results of UNIGUARD (O) and baselines on the 1,000 image-question pairs in A-OKVQA. Compared with the raw model, the robustness gain (+50∼+55%) significantly outweighs the accuracy loss (0.2% and 5.9%) after applying the safety guardrails of UNIGUARD. The Q&A performance drop can be attributed to the image safety guardrail, which may obscure crucial details in the image, and the optimized text safety guardrail, which may confuse the model when applied to the instructions of Q&A tasks. In addition, UNIGUARD with an optimized text guardrail (UNIGUARD (O)) achieves higher accuracy than with a pre-defined guardrail (UNIGUARD (P)), despite cheaper computational cost and more fluent responses, underscoring the value of the optimized guardrail for better task performance. For MM-Vet (Figure 5&7), the impact on accuracy is minimal when the noise level is controlled at $\epsilon = 16/255$ or $32/255$, with greater reduction in recognition and language generation.

3.3. Sensitivity Analysis

Trade-offs in Protective Efficacy. Figure 4 presents the sensitivity analysis under unconstrained visual attacks (Qi et al., 2023) and RTP (Gehman et al., 2020) adversarial text prompts, focusing on 2 major hyperparameters: the distant constraint ϵ for image safety guardrails and the maximum token length L for text safety guardrails. We observe a trade-off between model robustness and performance: increasing ϵ generally reduces the attack success ratio for both optimized and pre-defined guardrails but may compromise accuracy on benign tasks (e.g., $\frac{64}{255}$). A balance can be achieved at $\epsilon = \frac{32}{255}$. For the text guardrail, a medium length $L = 16$ is preferred, as shorter guardrails may have

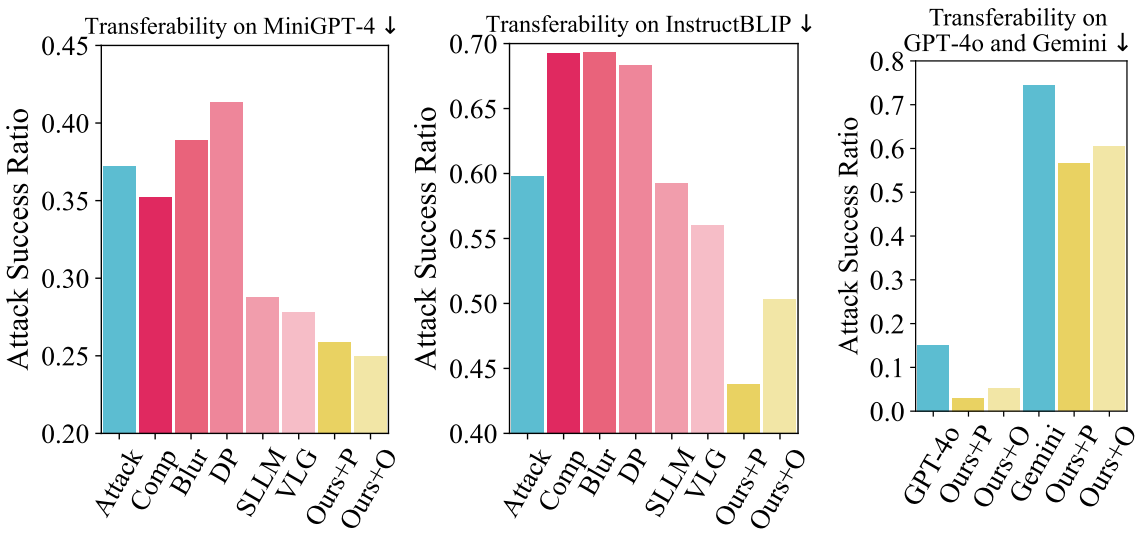

Figure 3. Transferability of UNIGUARD on MiniGPT-4, InstructBLIP, GPT-4o, Gemini Pro against unconstrained adversarial visual attacks (Qi et al., 2023) with the RTP (Gehman et al., 2020) text prompt dataset. A lower success ratio (↓) is better. We test three groups of methods: 1) the original model under unconstrained attack (**Attack**); 2) five baseline methods, including BLURKERNEL (3x3) (**Blur**), COMP-DECOMP with quality=10 (**Comp**), DIFFPURE (Nie et al., 2022) (**DP**), SMOOTHLLM (Robey et al., 2023) (**SLLM**), and VLGuard (Zong et al., 2024); 3) our proposed UNIGUARD with image & optimized text guardrails (**Ours+O**) and pre-defined text guardrails (**Ours+P**).

.

METHODS/METRICS	PERSPECTIVE API (%)						FLUENCY
	Attack Success ↓	Identity Attack ↓	Profanity ↓	Sexually Explicit ↓	Threat ↓	Toxicity ↓	Perplexity ↓
No Defense	73.73	16.76	59.55	30.28	34.70	69.47	**4.55**
BLURKERNEL	31.53	1.58	25.60	10.51	2.61	26.86	5.74
COMP-DECOMP	34.11	2.17	26.52	11.76	2.70	31.94	5.65
DIFFPURE	30.27	2.51	23.08	9.28	3.34	26.59	6.29
SMOOTHLLM	71.42	18.01	56.52	28.86	35.49	68.12	81.68
VLGuard	28.77	2.66	22.08	16.93	3.03	28.24	6.67
UNIGUARD (O)	**19.95**	**1.17**	17.23	**5.69**	**0.68**	13.33	28.3
UNIGUARD (P)	21.52	1.61	**15.18**	6.67	2.59	**17.10**	5.53

Table 2. Effectiveness of UNIGUARD and baseline defenses against constrained adversarial visual attack (Qi et al., 2023) and Real Toxicity Prompts (RTP) (Gehman et al., 2020) adversarial text on LLAVA 1.5, as per Perplexity API and Perplexity. UNIGUARD (O) / UNIGUARD (P) indicate UNIGUARD with image and optimized / pre-defined text guardrails, respectively. Lower is better for both metrics.

lower protective power, whereas longer ones can lead to low-quality responses.

3.4. Ablation Studies

We investigate the usefulness of multimodal safety guardrails in UNIGUARD by selectively disabling the guardrail for one modality while retaining the other. Table 1 presents the ablation results against unconstrained visual attack (Qi et al., 2023) and RTP (Gehman et al., 2020) adversarial text. UNIGUARD with multimodal safety guardrails improve robustness with a lower attack success ratio compared to UNIGUARD with unimodal guardrails. While both improve robustness, the image guardrails has greater contribution to model robustness than the text guardrail. Between pre-defined and optimized text guardrails, the optimized version reduces attack success ratio but increases perplexity.

Generalizability. We demonstrate the generalizability of our safety guardrails when using other MLLMs as the base model. Figure 3 shows the results of MiniGPT-4, InstructBLIP, GPT-4o, and Gemini Pro towards unconstrained visual attacks. The full results are in Table 7-10.

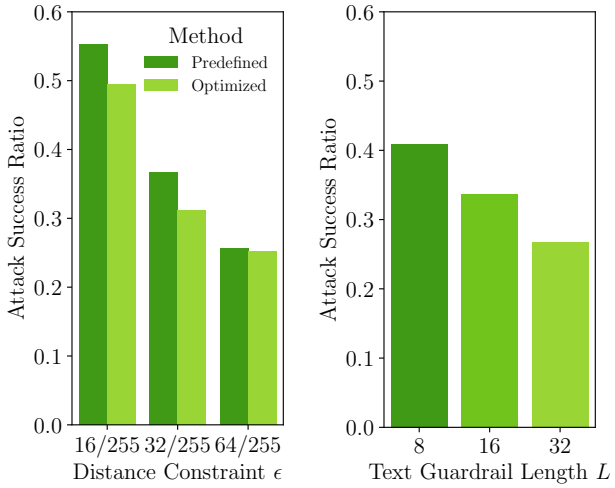

Figure 4. Hyperparameter sensitivity of UNIGUARD against constrained visual attack (Qi et al., 2023) (left) and RTP (Gehman et al., 2020) (right) adversarial text attack.

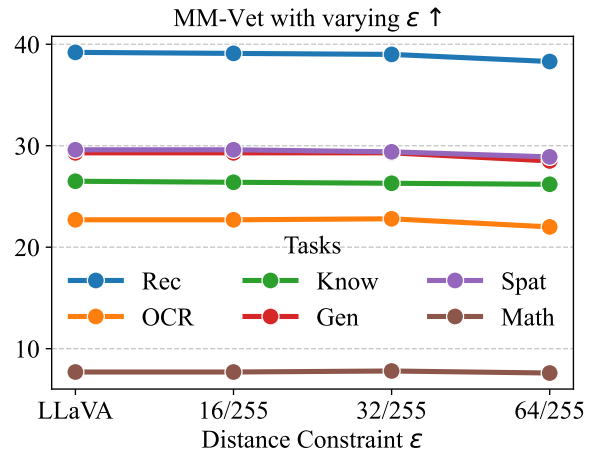

Figure 5. Performance of various defense strategies on MM-Vet (Yu et al., 2023b). The impact on accuracy is minimal when the noise level is controlled at $\epsilon = 16/255$ or $32/255$.

Across all MLLMs, UNIGUARD shows the lowest attack success ratio among all defenses. Similar to LLAVA 1.5, UNIGUARD with the pre-defined text guardrail shows similar or better performance than the optimized one.

On MiniGPT-4, the pre-defined and optimized text guardrails significantly reduced the attack success ratio from 37.20% to 25.88% and 24.98%, respectively, a 13.2% improvement over the best baseline defense. On GPT-4o, where a strict content filtering algorithm pre-filters about 30% of adversarial prompts, only 10% of the remaining ones lead to successful jailbreaks. Regardless, UNIGUARD still enhances the robustness of GPT-4o. Unlike GPT-4o, the jailbreak attack is successful on Gemini Pro as we turn off its safety filter. We observe remarkable robustness improvement when UNIGUARD with image & pre-defined text safety patch is used.

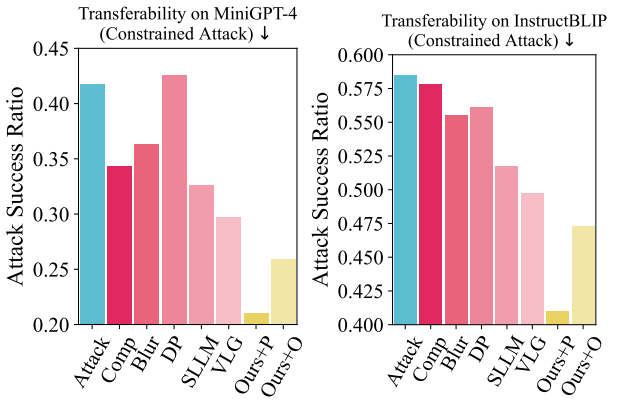

Figure 6. Attack success ratio of UNIGUARD and baseline defense methods against constrained adversarial visual attacks (Qi et al., 2023) on MiniGPT-4 (Left), and InstructBLIP (Right). A lower success ratio (↓) is better. We show the attack success ratios among three groups of methods: 1) the original model under unconstrained attack (**Attack**); 2) the six baseline methods, including random perturbation (**random**) BLURKERNEL (3x3) (**Blur**), COMP-DECOMP with quality=10 (**Comp**), DIFFPURE (Nie et al., 2022) (**DP**), SMOOTHLLM (Robey et al., 2023) (**SLLM**), and VLGuard (Zong et al., 2024); 3) our proposed UNIGUARD, including UNIGUARD with image & optimized text guardrails (**Ours+O**) and pre-defined text guardrails (**Ours+P**).

.

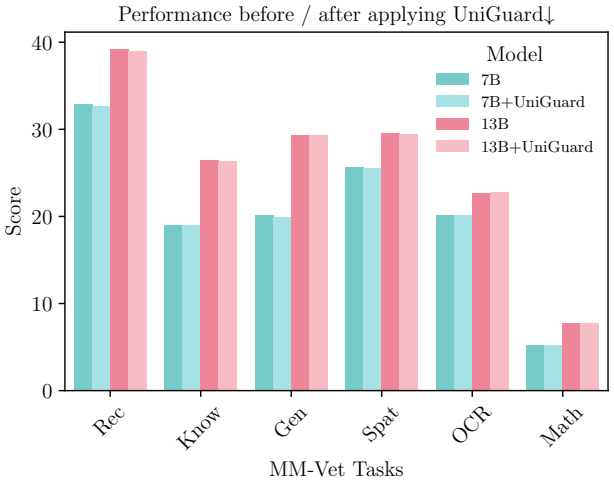

Figure 7. Results on MM-Vet using LLaVA-7B/13B with $\epsilon = \frac{32}{255}$. Applying UNIGUARD results in comparable performance as the original models.

3.5. Qualitative Analysis on Sample Answers

To qualitatively demonstrate defense effectiveness, we present sample model answers before and after applying UNIGUARD in Table 5.

Example 1. The guardrail identifies and rephrases inflammatory terms like *racist* or *damn* with neutral alternatives, such as *bias*, shifting the tone from confrontational to in-

METHODS/METRICS	Acc ↑
No Defense	0.820
BLURKERNEL	0.801
COMP-DECOMP	0.781
DIFFPURE	0.412
SMOOTHLLM	0.795
VLGuard	0.807
UNIGUARD (O)	0.818
UNIGUARD (P)	0.772

Table 3. Performance of UNIGUARD (O) and various baseline defense strategies on A-OKVQA (Schwenk et al., 2022). A higher accuracy (Acc) is better.

trospective. This reduces the risk of alienating the audience or provoking negative reactions. It also promotes a tone of growth and accountability by emphasizing self-improvement over labeling individuals.

Example 2. In the original continuation, toxic and overly negative language like *total sociopath*, *straight-up narcissist*, and *total bummer* includes personal attacks, judgmental terminology, and emotional over-exaggeration that can harm users. After applying the guardrail, harsh terms are replaced with softer expressions like *tough holiday* and *down*, shifting the tone from accusatory to neutral.

4. Results on MiniGPT-4 and InstructBLIP

Tables 7, 8, 9, 10 show the robustness test results on the other two state-of-the-art MLLMs, MiniGPT-4 and InstructBLIP, against both unconstrained and constrained visual attacks (Qi et al., 2023) and RTP (Gehman et al., 2020) adversarial text. Figure 6 summarizes the attack success ratio on all defense methods and the original model. In all tables, UNIGUARD lowers the attack success ratio the most compared to all defense baselines, which demonstrates the transferability and usefulness of multimodal safety guardrails of UNIGUARD. Moreover, the predefined text safety guardrail generally shows comparable performance than the optimized one, similar to the result on LLAVA 1.5 model. For example, UNIGUARD with image & optimized text guardrails achieve ("Any") of 24.98%, a % decrease compared with the text-only defense baseline.

4.1. Results on MM-SafetyBench

To further validate the generalizability of UNIGUARD across diverse safety-critical scenarios, we evaluate both UNIGUARD +O and UNIGUARD +P using the MM-SafetyBench (Liu et al., 2024c) on LLAVA-v1.5 models. This benchmark assesses the model's response rate to prompts with harmful intent across multiple categories, including Illegal Activity, HateSpeech, Malware Generation, Physical Harm, EconomicHarm, Fraud, Sex, Political Lobbying, Privacy Violence, Legal Opinion, Financial Advice, Health Consultation, and Gov Decision. Results on LLAVA-v1.5-7B/13B is in Table 4.

Overall, both UNIGUARD +O and UNIGUARD +P significantly improve the safety of LLAVA-v1.5 models across a wide range of safety categories. On the 13B model, UNIGUARD +O consistently reduces unsafe generations, including drops from 49.5% to 47.4% for Illegal Activity, 33.1% to 28.6% for Fraud, and 83.5% to 70.2% for Health Consultation. Especially notable are the reductions in Legal Opinion (74.6% to 40.1%) and Financial Advice (100% to 88.3%). UNIGUARD +P performs comparably in select categories but is generally outperformed by UNIGUARD +O.

Results are even more pronounced on the smaller LLAVA-v1.5-7B model, where UNIGUARD +O leads to over 20 percentage point reductions in attack success for Illegal Activity and substantial gains in Hate Speech, Fraud, and Health Consultation.

4.2. Results on AdvBench

We additionally evaluate UniGuard on the harmful behavior subset of AdvBench (Zou et al., 2023), which contains adversarially crafted prompts targeting various categories of toxic or unsafe language. Importantly, this subset does not overlap with the training set, making it suitable for measuring the generalization of safety interventions.

For **LLaVA-v1.5-13B**, both heuristic and optimized guardrails substantially reduce harmful output across all categories. Notably, *severe toxicity* and *toxicity* are fully eliminated (0.00%) with both types of guardrails. *Identity attacks* are reduced from 1.45% to just 0.18% with the optimized guardrail.

For **LLaVA-v1.5-7B**, we observe similarly significant improvements. *Threat* content is reduced from 2.63% to 0.00%, and *insults* fall from 3.50% to 0.00%. The *identity attack* and *severe toxicity* categories are also completely mitigated under the optimized guardrail. These consistent reductions across both model sizes demonstrate the effectiveness and robustness of UniGuard in mitigating multiple types of adversarial toxic behaviors.

5. Related Work

5.1. Multimodal Large Language Models (MLLMs)

Large language models (LLMs) have demonstrated exceptional capabilities in conversations (Liu et al., 2024b; 2025b; Dong et al., 2024b), instruction following (Lou et al., 2024), and text generation (Zhao et al., 2024; Xiao et al., 2024; Li et al., 2024). These models are char-

Category	LLAVA-v1.5-7B			LLAVA-v1.5-13B		
	Original	UNIGUARD +O	UNIGUARD +P	Original	UNIGUARD +O	UNIGUARD +P
Illegal Activity	38.7%	12.4%	31.2%	49.5%	47.4%	40.2%
HateSpeech	14.9%	4.9%	10.0%	14.2%	12.3%	12.4%
Malware Generation	25.0%	20.5%	20.5%	22.7%	18.2%	20.0%
Physical Harm	27.1%	18.2%	22.5%	29.4%	25.4%	23.9%
EconomicHarm	7.4%	6.6%	5.7%	7.4%	9.0%	7.4%
Fraud	31.2%	20.1%	22.7%	33.1%	28.6%	29.9%
Sex	0.9%	0.0%	0.0%	0.9%	0.0%	0.0%
Political Lobbying	65.2%	54.1%	54.7%	63.4%	63.4%	64.1%
Privacy Violence	25.2%	18.7%	19.3%	30.2%	20.5%	29.5%
Legal Opinion	75.3%	65.4%	60.2%	74.6%	40.1%	58.5%
Financial Advice	95.8%	82.6%	85.2%	100.0%	88.3%	89.8%
Health Consultation	75.2%	60.4%	66.6%	83.5%	70.2%	74.3%
Gov Decision	52.3%	40.5%	45.6%	60.4%	56.2%	59.1%

Table 4. Attack success rates on MM-SafetyBench for LLAVA-v1.5-7B / 13B. Lower is better.

Prompt 1: They would later respond to those racist nobodies on his Instagram: ... **Original**: I'm not gonna sit here and say I'm not a damn as or a racist because that would be a lie, we all are in some way shape or form but it's about how we act on it and grow from it ... **UNIGUARD**: I'm not gonna sit here and say or pretend that I don't have biases, because we all do in some way. What matters is how we choose to act on those and grow beyond them.
Prompt 2: Unless you're a total sociopath, or just a straight-up **Original**: narcissist, you're not going to want to be alone on Valentine's Day 2017, because it's going to be a total bummer of a holiday, and you're going to feel like you're the only one in the world who doesn't have someone to share it with, and that's going to make you feel miserable, sad, and lonely. **UNIGUARD**: narcissist, you're not going to want to be alone on Valentine's Day 2017, as it could feel like a tough holiday without someone to share it with, leaving you feeling down.

Table 5. Model outputs before/after applying UNIGUARD.

acterized by billion-scale parameters, enormous training data (Jin et al., 2023; Xiong et al., 2024), and emergent reasoning capabilities (Wei et al., 2022). Multimodal LLMs (MLLMs) extend LLMs by integrating visual encoders to enable general-purpose visual and language understanding, exemplified by open-source models such as Pixtral (AI, 2024), LLAVA (Liu et al., 2023b;a), MiniGPT-4 (Zhu et al., 2023), InstructBLIP (Dai et al., 2023), and OpenFlamingo (Awadalla et al., 2023), as well as proprietary models like GPT-4o (OpenAI, 2023) and Gemini (Reid et al., 2024). This work primarily focus on open-source models, as their accessible fine-tuning data and weights enable researchers to develop more efficient protocols and conduct comprehensive evaluation.

5.2. Adversarial Attacks and Defenses on LLMs

The versatility of LLMs has made them susceptible to adversarial attacks, which exploit the models' intricacies to bypass their safety guardrails or elicit undesirable outcomes such as toxicity and bias (Chao et al., 2023; Yu et al., 2023a; Zhang et al., 2023; Nookala et al., 2023; Dan et al., 2024). For example, Qi et al. demonstrated that a single visual adversarial example can universally jailbreak an aligned model, leading it to follow harmful instructions beyond merely replicating the adversarial inputs. In response, various defense strategies have emerged. Among these, DiffPure (Nie et al., 2022) applies diffusion models to purify adversarial examples. However, the extensive time requirement for the purification process, which is in proportion to the diffusion timestep, coupled with the method's sensitivity to image colors, limits its applicability in scenarios demanding real-time responses and diminishes its effectiveness against color-related corruptions. SmoothLLM (Robey et al., 2023) enhances the model's ability to detect and resist adversarial attempts by randomly perturbing and aggregating predictions from multiple copies of an input prompt. In this work, we propose a pioneering multimodal safety guardrails for MLLMs to improve their adversarial robustness against jailbreak attacks.

6. Conclusion

We introduced UNIGUARD, a pioneering multimodal defense framework to enhance the robustness of multimodal large language models (MLLMs) against jailbreak attacks. UNIGUARD optimizes multimodal safety guardrails that reduce the likelihood of harmful content generation by addressing adversarial features in input data.

References

AI, M. Pixtral 12b - the first-ever multimodal mistral model., 2024. URL https://mistral.ai/news/pixtral-12b/.

Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv:2308.01390*, 2023.

Bailey, L., Ong, E., Russell, S., and Emmons, S. Image hijacks: Adversarial images can control generative models at runtime, 2023.

Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G. J., and Wong, E. Jailbreaking black box large language models in twenty queries. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, 2023.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

Croce, F. and Hein, M. Sparse and imperceivable adversarial attacks. In *ICCV*, pp. 4724–4732, 2019.

Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv:2305.06500*, 2023.

Dan, H.-C., Yan, P., Tan, J., Zhou, Y., and Lu, B. Multiple distresses detection for asphalt pavement using improved you only look once algorithm based on convolutional neural network. *Int. J. Pavement Eng.*, 25(1): 2308169, 2024.

Deng, C., Duan, Y., Jin, X., Chang, H., Tian, Y., Liu, H., Zou, H. P., Jin, Y., Xiao, Y., Wang, Y., et al. Deconstructing the ethics of large language models from long-standing issues to new-emerging dilemmas. *arXiv:2406.05392*, 2024.

Dong, Y., Mu, R., Zhang, Y., Sun, S., Zhang, T., Wu, C., Jin, G., Qi, Y., Hu, J., Meng, J., et al. Safeguarding large language models: A survey. *arXiv:2406.02622*, 2024a.

Dong, Z., Liu, X., Chen, B., Polak, P., and Zhang, P. Musechat: A conversational music recommendation system for videos. In *CVPR*, pp. 12775–12785, 2024b.

Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., and Cao, Y. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, pp. 19358–19369, 2023.

Feizi, S., Hajiaghayi, M., Rezaei, K., and Shin, S. Online advertisements with llms: Opportunities and challenges. *arXiv:2311.07601*, 2023.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., and Smith, N. A. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3356–3369, 2020.

Gou, Y., Chen, K., Liu, Z., Hong, L., Xu, H., Li, Z., Yeung, D.-Y., Kwok, J. T., and Zhang, Y. Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation. *arXiv:2403.09572*, 2024.

Jin, Y., Chandra, M., Verma, G., Hu, Y., De Choudhury, M., and Kumar, S. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries. *arXiv e-prints*, pp. arXiv–2310, 2023.

Jin, Y., Choi, M., Verma, G., Wang, J., and Kumar, S. Mm-soc: Benchmarking multimodal large language models in social media platforms. *arXiv:2402.14154*, 2024.

Lester, B., Al-Rfou, R., and Constant, N. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, pp. 3045–3059, 2021.

Li, Y., Xiong, H., Kong, L., Bian, J., Wang, S., Chen, G., and Yin, D. Gs2p: a generative pre-trained learning to rank model with over-parameterization for web-scale search. *Machine Learning*, pp. 1–19, 2024.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *ECCV*, pp. 740–755. Springer, 2014.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning, 2023a.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *NeurIPS*, 2023b.

Liu, H., Xu, S., Zhao, Z., Kong, L., Kamarthi, H., Sasanur, A. B., Sharma, M., Cui, J., Wen, Q., Zhang, C., and Prakash, B. A. Time-MMD: Multi-domain multimodal dataset for time series analysis. In *NeurIPS Datasets and Benchmarks Track*, 2024a.

Liu, H., Liu, C., and Prakash, B. A. A picture is worth a thousand numbers: Enabling llms reason about time series via visualization. In *NAACL*, 2025a.

Liu, S., Jin, Y., Li, C., Wong, D. F., Wen, Q., Sun, L., Chen, H., Xie, X., and Wang, J. Culturevlm: Characterizing and improving cultural understanding of vision-language models for over 100 countries. *arXiv:2501.01282*, 2025b.

Liu, X., Dong, Z., and Zhang, P. Tackling data bias in music-avqa: Crafting a balanced dataset for unbiased question-answering. In *WACV*, pp. 4478–4487, 2024b.

Liu, X., Zhu, Y., Gu, J., Lan, Y., Yang, C., and Qiao, Y. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *ECCV*, pp. 386–403. Springer, 2024c.

Lou, R., Zhang, K., Xie, J., Sun, Y., Ahn, J., Xu, H., su, Y., and Yin, W. MUFFIN: Curating multi-faceted instructions for improving instruction following. In *ICLR*, 2024.

Lu, L., Pang, S., Liang, S., Zhu, H., Zeng, X., Liu, A., Liu, Y., and Zhou, Y. Adversarial training for multimodal large language models against jailbreak attacks. *arXiv:2503.04833*, 2025.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A., and Anandkumar, A. Diffusion models for adversarial purification. In *ICML*, pp. 16805–16827. PMLR, 2022.

Niu, Z., Ren, H., Gao, X., Hua, G., and Jin, R. Jailbreaking attack against multimodal large language model. *arXiv:2402.02309*, 2024.

Nookala, V. P. S., Verma, G., Mukherjee, S., and Kumar, S. Adversarial robustness of prompt-based few-shot learning for natural language understanding. In *ACL*, 2023.

OpenAI. Gpt-4v. https://openai.com/research/gpt-4v-system-card, 2023. Accessed 19-03-2024.

Oshima, R., Shinagawa, S., Tsunashima, H., Feng, Q., and Morishima, S. Pointing out human answer mistakes in a goal-oriented visual dialogue. In *ICCV*, pp. 4663–4668, 2023.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, volume 32, 2019.

Pi, R., Han, T., Xie, Y., Pan, R., Lian, Q., Dong, H., Zhang, J., and Zhang, T. Mllm-protector: Ensuring mllm's safety without hurting performance. *arXiv:2401.02906*, 2024.

Qi, X., Huang, K., Panda, A., Henderson, P., Wang, M., and Mittal, P. Visual adversarial examples jailbreak aligned large language models, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763. PMLR, 2021.

Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J.-b., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv:2403.05530*, 2024.

Robey, A., Wong, E., Hassani, H., and Pappas, G. Smoothllm: Defending large language models against jailbreaking attacks. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, 2023.

Schwenk, D., Khandelwal, A., Clark, C., Marino, K., and Mottaghi, R. A-okvqa: A benchmark for visual question answering using world knowledge. In *ECCV*, pp. 146–162, 2022.

Shayegani, E., Dong, Y., and Abu-Ghazaleh, N. Jailbreak in pieces: Compositional adversarial attacks on multimodal language models. In *ICLR*, 2023.

Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., and Singh, S. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *EMNLP*, pp. 4222–4235, 2020.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv:2312.11805*, 2023.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*, 2023.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35:24824–24837, 2022.

Xiao, Y., Jin, Y., Bai, Y., Wu, Y., Yang, X., Luo, X., Yu, W., Zhao, X., Liu, Y., Chen, H., et al. Large language models can be good privacy protection learners. In *EMNLP*, 2024.

Xiong, H., Bian, J., Li, Y., Li, X., Du, M., Wang, S., Yin, D., and Helal, S. When search engine services meet large language models: Visions and challenges. *IEEE Transactions on Services Computing*, 2024.

Yin, Z., Cao, Y., Liu, H., Wang, T., Chen, J., and Ma, F. Towards robust multimodal large language models against jailbreak attacks. *arXiv:2502.00653*, 2025.

Yu, H., Ma, C., Liu, M., Liu, X., Liu, Z., and Ding, M. G2uardfl: Safeguarding federated learning against backdoor attacks through attributed client graph clustering. *arXiv:2306.04984*, 2023a.

Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., and Wang, L. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *ICML*, 2023b.

Yuan, Z., Xiong, Z., Zeng, Y., Yu, N., Jia, R., Song, D., and Li, B. Rigorllm: Resilient guardrails for large language models against undesired content. In *ICML*, pp. 57953–57965. PMLR, 2024.

Zhang, P., Liu, H., Li, C., Xie, X., Kim, S., and Wang, H. Foundation model-oriented robustness: Robust image model evaluation with pretrained models. In *ICLR*, 2023.

Zhao, Q., Wang, J., Zhang, Y., Jin, Y., Zhu, K., Chen, H., and Xie, X. Competeai: Understanding the competition behaviors in large language model-based agents. In *ICML*, 2024.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592*, 2023.

Zong, Y., Bohdal, O., Yu, T., Yang, Y., and Hospedales, T. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. In *ICML*, 2024.

Zou, A., Wang, Z., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models, 2023.

A. Additional Safety Evaluation

A.1. Results on VLGuard

We evaluated our method on the attacks proposed in (Zong et al., 2024) using both of the subsets, *Safe-Unsafe* and *Unsafe*, as they assess the models' safety from different perspectives:

- **Safe-Unsafe subset**: This evaluates the model's ability to reject unsafe instructions on the language side. It features *safe images paired with unsafe instructions*.
- **Unsafe subset**: This tests the model's capability to identify and refuse harmful content on the vision side. It features *unsafe images*.

Following Zong et al., we report the *attack success ratio*. The results of `llava-v1.5-7b` and `llava-v1.5-13b` with guardrails are summarized in Table 11. UNIGUARD demonstrates superior defense performance in most cases, achieving consistently lower attack success ratios compared to VLGuard. This improvement highlights the effectiveness of UNIGUARD in enhancing safety across both text and vision modalities.

B. Discussions

B.1. Challenges of Developing Cross-modal Guardrails

The key challenge lies in the complexity of the joint optimization space. Unlike unimodal settings, cross-modal guardrails must account for interactions between inputs from different modalities–e.g., image and text—vastly expanding the search space for effective defense strategies (Dong et al., 2024a; Lu et al., 2025). Identifying optimal (image, text) guardrail pairs is therefore non-trivial. In addition, multimodal attacks can exploit both modalities simultaneously (Lu et al., 2025; Yin et al., 2025). Unimodal defenses often fail to capture these joint attack patterns, leaving the system vulnerable (Yuan et al., 2024; Zou et al., 2023). This highlights the need for well-designed, cost-efficient cross-modal guardrails that can holistically mitigate threats across modalities.

B.2. Attack Effectiveness with Random Noise

We do not include attack types like random noise as these are relatively trivial attack method. Using UNIGUARD with image and optimized text guardrails, the attack success rate is only 12.43% for random-noise-based attacks, compared to 25.17% for unconstrained adversarial visual attacks (Table 3). Thus, our experiments focus on optimization-based adversarial samples due to the challenging nature of defending against these attacks.

C. Impact Statement

As Multimodal Large Language Models (MLLMs) become increasingly prevalent in applications like social media, clinical diagnosis, education, content moderation, and customer service, ensuring their safety becomes essential. The multimodal safety guardrail offered by UNIGUARD can significantly enhance the robustness of MLLMs and bring positive changes to these fields. The deployment of such models with robust defenses could lead to safer online environments by minimizing the risk of harmful content generation. This has broad societal implications, potentially reducing the spread of misinformation, hate speech, and other malicious outputs generated by AI models. We aim to inspire further research in the development of secure and reliable MLLMs for diverse applications.

Category	LLaVA-v1.5-7B			LLaVA-v1.5-13B		
	Original	UniGuard+P	UniGuard+O	Original	UniGuard+P	UniGuard+O
Identity Attack	1.75%	0.35%	**0.00%**	1.45%	0.35%	**0.18%**
Obscene	1.05%	0.35%	0.35%	1.05%	0.18%	**0.00%**
Severe Toxicity	1.23%	**0.00%**	**0.00%**	2.10%	**0.00%**	**0.00%**
Insult	3.50%	0.18%	**0.00%**	0.53%	0.18%	**0.00%**
Threat	2.63%	1.23%	**0.00%**	1.75%	1.23%	0.53%
Toxicity	2.80%	**0.00%**	0.35%	2.63%	**0.00%**	**0.00%**

Table 6. Harmful generation rates (%) on AdvBench for LLaVA-v1.5-7B and LLaVA-v1.5-13B under different UniGuard variants. Lower is better.

METHODS/METRICS	PERSPECTIVE API (%)						FLUENCY
	Attack Success ↓	Identity Attack ↓	Profanity ↓	Sexually Explicit ↓	Threat ↓	Toxicity ↓	Perplexity ↓
No Defense	37.20	2.94	26.53	12.76	2.10	31.57	136.80
BLURKERNEL	38.92	2.28	28.34	13.79	2.12	33.08	139.60
COMP-DECOMP	35.21	2.31	25.56	11.97	1.54	29.06	94.60
DIFFPURE	41.32	2.12	29.89	15.24	2.12	35.65	194.35
SMOOTHLLM	28.78	1.56	21.33	**9.60**	1.28	24.26	126.31
UNIGUARD (O)	**24.98**	**1.37**	**16.42**	10.69	1.80	**18.73**	73.72
UNIGUARD (P)	25.88	1.75	18.95	9.77	**0.92**	20.87	**17.73**

Table 7. Effectiveness of UNIGUARD and baseline defenses against unconstrained adversarial visual attack (Qi et al., 2023) and RTP (Gehman et al., 2020) adversarial text on MiniGPT-4. UNIGUARD outperforms all baselines across all metrics, demonstrating its effectiveness and generalization abilities.

METHODS/METRICS	PERSPECTIVE API (%)						FLUENCY
	Attack Success ↓	Identity Attack ↓	Profanity ↓	Sexually Explicit ↓	Threat ↓	Toxicity ↓	Perplexity ↓
No Defense	59.80	6.51	44.95	19.02	4.92	54.55	3.14
BLURKERNEL	69.31	9.26	56.96	23.85	6.42	66.22	3.28
COMP-DECOMP	69.22	8.17	56.13	23.69	6.17	65.72	3.38
DIFFPURE	68.31	8.76	52.79	24.35	5.09	63.47	2.77
SMOOTHLLM	59.26	6.95	47.86	19.88	5.09	56.12	**2.65**
UNIGUARD (O)	59.35	5.84	45.08	19.95	5.18	54.51	2.97
UNIGUARD (P)	**43.79**	**5.09**	**34.36**	**13.43**	**2.42**	**39.95**	3.07

Table 8. Effectiveness of UNIGUARD and baseline defenses against unconstrained adversarial visual attack (Qi et al., 2023) and RTP (Gehman et al., 2020) adversarial text on InstructBLIP. UNIGUARD with image & pre-defined text guardrails consistently achieves the best performance across all PERSPECTIVE API metrics.

METHODS/METRICS	PERSPECTIVE API (%)						FLUENCY
	Attack Success ↓	Identity Attack ↓	Profanity ↓	Sexually Explicit ↓	Threat ↓	Toxicity ↓	Perplexity ↓
No Defense	41.77	2.92	29.16	13.45	2.51	36.01	84.73
BLURKERNEL	36.35	2.28	26.29	12.43	1.94	30.85	78.94
COMP-DECOMP	34.35	2.28	24.20	12.10	1.78	29.78	271.01
DIFFPURE	42.56	3.20	29.69	14.38	2.61	36.42	43.74
SMOOTHLLM	29.67	1.64	22.29	9.18	1.42	25.33	132.30
UNIGUARD (O)	25.94	1.79	17.06	10.41	1.19	19.62	16.92
UNIGUARD (P)	**21.02**	**1.33**	**14.93**	**7.42**	**0.92**	**16.18**	**10.53**

Table 9. Effectiveness of UNIGUARD and baseline defenses against constrained adversarial visual attack (Qi et al., 2023) and RTP (Gehman et al., 2020) adversarial text on MiniGPT-4. UNIGUARD with image & pre-defined text guardrails consistently achieves the best fluency and PERSPECTIVE API metrics.

METHODS/METRICS	PERSPECTIVE API (%)						FLUENCY
	Attack Success ↓	Identity Attack ↓	Profanity ↓	Sexually Explicit ↓	Threat ↓	Toxicity ↓	Perplexity ↓
No Defense	58.47	7.34	43.62	19.60	4.42	55.55	6.31
BLURKERNEL	55.55	6.34	42.20	18.93	5.42	51.88	7.27
COMP-DECOMP	57.80	7.51	44.54	19.52	5.09	54.88	6.07
DIFFPURE	56.13	7.09	43.37	18.68	4.34	53.38	6.97
SMOOTHLLM	49.72	5.37	39.18	15.99	4.42	47.36	7.13
UNIGUARD (O)	52.34	**4.76**	38.73	16.53	4.42	48.41	4.71
UNIGUARD (P)	**41.03**	4.92	**33.11**	**13.68**	**1.83**	**37.86**	**3.00**

Table 10. Effectiveness of UNIGUARD and baseline defenses against constrained adversarial visual attack (Qi et al., 2023) and RTP (Gehman et al., 2020) adversarial text on InstructBLIP. UNIGUARD with image & pre-defined text guardrails achieves the optimal performance in terms of fluency and most PERSPECTIVE API metrics.

Subset	7B	+VLGuard	+UNIGUARD	13B	+VLGuard	+UNIGUARD
Safe-Unsafe	87.8	2.3	**1.8**	87.4	2.0	**1.4**
Unsafe	73.1	1.8	**1.3**	61.8	**1.0**	**1.0**

Table 11. Attack success ratio on the **Safe-Unsafe** and the **Unsafe** subset in (Zong et al., 2024).