SmoothGuard: Defending Multimodal Large Language Models with Noise Perturbation and Clustering Aggregation

Guangzhi Su[1,*], Shuchang Huang[2,*], Yutong Ke[1], Zhuohang Liu[1], Long Qian[1], Kaizhu Huang[1,†]
[1] Division of Natural and Applied Sciences, Duke Kunshan University, Suzhou, China
guangzhi.su, yutong.ke, zhuohang.liu, long.qian, kaizhu.huang@dukekunshan.edu.cn
[2] Independent Researcher
shuchanghuang2001@gmail.com

***Abstract*—Multimodal large language models (MLLMs) have achieved impressive performance across diverse tasks by jointly reasoning over textual and visual inputs. Despite their success, these models remain highly vulnerable to adversarial manipulations, raising concerns about their safety and reliability in deployment. In this work, we first generalize an approach for generating adversarial images within the HuggingFace ecosystem and then introduce SmoothGuard, a lightweight and model-agnostic defense framework that enhances the robustness of MLLMs through randomized noise injection and clustering-based prediction aggregation. Our method perturbs continuous modalities (e.g., images and audio) with Gaussian noise, generates multiple candidate outputs, and applies embedding-based clustering to filter out adversarially influenced predictions. The final answer is selected from the majority cluster, ensuring stable responses even under malicious perturbations. Extensive experiments on POPE, LLaVA-Bench (In-the-Wild), and MM-SafetyBench demonstrate that SmoothGuard improves resilience to adversarial attacks while maintaining competitive utility. Ablation studies further identify an optimal noise range (0.1–0.2) that balances robustness and utility. These findings establish SmoothGuard as a practical step toward secure and reliable multimodal reasoning, with future work extending the framework to audio-focused benchmarks to validate its generality across modalities. Code will be publicly available soon on Github.**

***Index Terms*—Multimodal large language models, adversarial defense, model robustness, trustworthy AI**

I. Introduction

In recent years, multi-modal large language models (MLLMs) [1]–[5] have surged in popularity, driven by their capacity to assimilate and analyze data across text and visual modalities. These models find applications in diverse domains, including content generation, question answering, and visual-language reasoning. Their sophisticated ability to synchronize multimodal inputs enables them to deliver coherent and precise responses, showcasing significant advancements in artificial intelligence technologies [6].

Despite their impressive capabilities, MLLMs are not immune to adversarial attacks that hurt their functionality and safety [1], [7]. Adversarial inputs were designed deliberately by data manipulations, which pose a serious threat as they can induce significant misclassifications and trigger potentially harmful responses. Current adversarial attacks on MLLMs are mostly targeted on continuous modality like image and audio by adding designed noise to get close and cross the decision boundary. These threats underscore the urgent need for robust defenses to facilitate the safe deployment of MLLMs in practical scenarios.

To this end, various strategies [8], [9]have been proposed to provide defenses for both large language models (LLMs) and MLLMs. Innovations such as SMOOTHLLM and patched visual prompt defenses represent notable attempts to safeguard these models. However, these measures often fall short when confronted with sophisticated, multi-modal adversarial strategies. What's more, in resource constrained circumstances, finetuning a model with a large designed dataset might be infeasible, which calls for a method that can neutralize the malicious prompt during inference stage, and a plug-and-play method that can be applied to different models directly.

To address these issues, we propose a defense strategy based on randomized noise injection and semantic clustering method for selecting the optimal answer. Unlike detector-based or heuristic defenses, our method leverages multiple perturbed inputs and aggregates their predictions through clustering, filtering out adversarially influenced responses while preserving stable ones. This design provides robustness against subtle, hard-to-detect attacks without requiring model retraining or architectural modifications. Initial experiments on vision–language benchmarks show that the method reduces adversarial risk while retaining competitive utility, with ablation studies identifying a practical range of noise levels that balances these objectives. Our work thus contributes a lightweight, model-agnostic approach to enhancing the reliability of MLLMs in adversarial settings.

This paper is structured as follows: Section 2 reviews related work in the realm of adversarial attacks and defenses for MLLMs. Section 3 explains our proposed methodology in detail. Section 4 describes our experimental framework and discusses the results. Section 5 concludes with a reflection on our findings and suggests directions for future research.

*Equal contribution. ‡Corresponding author.

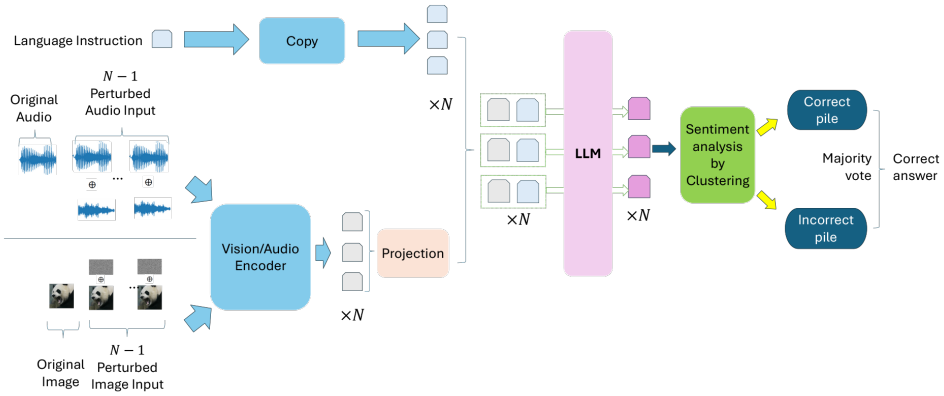

Fig. 1. Overall pipeline of our randomized smoothing defense for MLLMs. Perturbed audio and image inputs are processed through the encoder and projection, producing N responses via the language model. Outputs are clustered and sentiment-aware majority voting is applied to select the final robust answer.

II. Related Work

A. Adversarial Attacks on Multimodal Models

The attack surface of multimodal MLLMs has expanded beyond unimodal paradigms [10]–[12], with cross-modal vulnerability propagation emerging as a critical threat vector. Crucially, multimodal jailbreaking exploits safety mechanisms via perturbation vectors [13], [14] and in-context learning backdoors [15], creating attack surfaces that transcend traditional unimodal paradigms. Contemporary research demonstrates that imperceptible perturbations can induce cascading failures across modalities: multimodal systems exhibit significant degradation in safety performance under semantically constrained adversarial attacks, where signal-level distortions propagate vulnerabilities between audio and textual domains [16], while audio-language systems exhibit alarming susceptibility to jailbreak attacks. Chen et al. [17] achieve asynchrony, universality, stealthiness, and over-the-air robustness at the same time via adversarial suffix injections to benign audio carriers. Ma et al. [18] compromise SpeechGPT through discrete token manipulation in HuBERT embeddings. These vulnerabilities are systematically quantified in the AJailBench benchmark [19], reporting a high attack success rate (ASR) across seven state-of-the-art models through Bayesian-optimized perturbations targeting ten prohibited categories.

B. Defensive Techniques in Single and Multimodal Models

Defensive strategies for single-modality models provide a foundation for protecting MLLMs. [20], [21] Unimodal foundations include randomized smoothing for certified vision robustness [22], ensemble methods preserving text integrity under synonym substitution attacks [23], and Time-Domain Noise Flooding (TDNF), reducing audio jailbreak ASR [24]. For multimodal scenarios, SPIRIT's neuron-level patching achieves a 99% defense success rate (DSR) through activation modification [25], while hybrid purification frameworks integrate input transformation with feature sanitization [15], [26]. Persistent limitations remain in mitigating real-time audio adversarial inputs [17] and certifying robustness against cross-modal perturbation transfer, particularly for safety-critical deployments requiring low-latency responses.

Our method extends these defense strategies by applying random sampling and aggregation techniques to image and audio inputs, improving the adversarial robustness of MLLMs. While previous work has focused on single-modality defenses or simplified approaches to multimodal models, our contribution offers a unified defense that covers both modalities. A novel clustering-based sentiment analysis method then enables extension to a much wider range of adversarial attack tasks.

III. Proposed Method

A. Overview

We propose a robust defense method for MLLMs as shown in Fig. 1, which involves injecting noise to continuous modalities (e.g. image/audio) to safeguard against adversarial attacks. Our strategy is to apply a diverse array of perturbations to the image and then aggregate the model outputs to generate a consolidated, robust prediction. This approach is designed to defend against sophisticated adversarial attacks by sampling a variety of perturbations and clustering the resulting outputs. This clustering helps to filter out outputs that are likely to have been influenced by adversarial inputs, thus ensuring a more reliable and secure response.

Additionally, we employ sentiment analysis based majority voting as a key component of our defense mechanism. This method allows us to evaluate the consensus among the different perturbed outputs, effectively defending against both task-specific and jailbreaking attacks. By doing so, we ensure that our model's responses remain dependable even under malicious conditions.

The effectiveness of our approach is further enhanced by integrating feedback loops from the sentiment analysis process, which help to continually refine the perturbation and clustering techniques. This adaptive mechanism not only improves the resilience of our MLLMs over time but also maintains high accuracy and performance across a range of tasks. By

providing a robust framework for handling adversarial inputs in multimodal scenarios, our method stands as a significant advancement in the field of AI security.

B. Random Sampling of Perturbations

Let x_{img}, x_{text}, and x_{audio} represent the clean image, text, and audio inputs, respectively. To improve robustness, we generate perturbed versions of the continuous modalities by introducing random noise. We keep text unchanged to preserve semantic integrity after observing a severe performance drop with changed text.

For the image input, Gaussian perturbations with variance σ^2_{img} are applied:

$$\delta_{\text{img}} \sim \mathcal{N}(0, \sigma^2_{\text{img}} I), \quad \tilde{x}_{\text{img}} = x_{\text{img}} + \delta_{\text{img}}. \tag{1}$$

For the audio input, we apply a similar perturbation at the waveform or spectrogram level:

$$\delta_{\text{audio}} \sim \mathcal{N}(0, \sigma^2_{\text{audio}} I), \quad \tilde{x}_{\text{audio}} = x_{\text{audio}} + \delta_{\text{audio}}. \tag{2}$$

The text input is used in its original form:

$$\tilde{x}_{\text{text}} = x_{\text{text}}. \tag{3}$$

Thus, for each sampling step, we obtain randomized variations of the image and audio inputs, while maintaining consistent textual guidance.

C. Aggregation of Predictions via Clustering

Once we obtain perturbed versions of the image and audio inputs, we also include one unperturbed copy of the original input to preserve utility. Each input pair $(\tilde{x}^{(i)}_{\text{img}}, \tilde{x}^{(i)}_{\text{audio}}, x_{\text{text}})$ is passed through the MLLM to generate a prediction:

$$y^{(i)}_{\text{pred}} = f(\tilde{x}^{(i)}_{\text{img}}, \tilde{x}^{(i)}_{\text{audio}}, x_{\text{text}}) \tag{4}$$

where $i \in \{1, 2, \ldots, N\}$ indexes the perturbed samples, and an additional sample corresponds to the original unperturbed input.

As shwon in Fig. 2, to aggregate predictions, we embed the outputs using a *RoBERTa-base* encoder, producing a fixed-dimensional representation $\mathbf{e}_i$ for each prediction. [27] We then apply *k-means clustering* with $k = 2$:

$$\text{clusters} = \text{k-means}(\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N\}, k = 2). \tag{5}$$

Intuitively, robust models tend to embed semantically-equivalent inputs into tight, well-separated clusters, while adversarial perturbations tend to push corrupted inputs away from the clean cluster into different regions in the representation space. This effect has been observed in other researchers' papers related to the "clustering effect" in robust networks) [28]. In practice, we often observe a dominant clean cluster covering most of the perturbed samples, with the remainder forming a distinct minority cluster corresponding to adversarial deviations. This motivates our use of k = 2 for clustering: one larger cluster for clean and stable predictions, and the other for adversarial ones.

We then select the larger cluster for next step process. The centroid of this cluster is computed, and the prediction that is the closest to the centroid was chosen as the representative answer for stability:

$$\hat{y} = \arg \max_{\mathbf{e} \in \text{cluster}} \cos(\mathbf{e}, \text{centroid}), \tag{6}$$

where $\cos(\mathbf{e}, \text{centroid})$ denotes the cosine similarity.

Including the unperturbed input ensures that, under benign conditions, the final prediction remains close to the model's original utility. In the presence of adversarial attacks, the use of multiple perturbed copies prevents a single corrupted prediction from dominating the outcome, thereby stabilizing the defense.

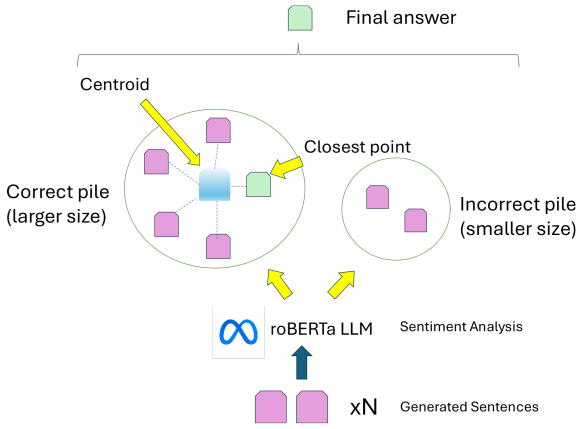

Fig. 2. Clustering-based aggregation of generated sentences.

IV. Experiment

A. Experimental Setup

All experiments were implemented using the HuggingFace pipeline, which allows for straightforward adaptation to different MLLMs. Our study focuses primarily on the Qwen2.5-VL-7B and LLaVA-1.5-7B models, with experiments organized into two parts: robustness evaluation and utility evaluation. The robustness evaluation measures defense performance under adversarially perturbed inputs, while the utility evaluation examines performance under benign conditions. For the latter, we consider two representative tasks: (1) multiple-choice question answering, and (2) free-form sentence-based question answering. Finally, we conduct an ablation study to identify the optimal range of noise levels, aiming to balance robustness improvements with minimal utility degradation.

B. Robustness Analysis

Benchmark and attack. For robustness analysis, we adopt MM-SafetyBench as the benchmark. [29] This benchmark comprises 13 scenarios and contains over 5,000 text–image pairs. In our evaluation, however, we replace the dataset images with a single universal adversarial image trained in advance, as this aligns with our target adversarial attack method and is generally more effective than standard attacks. Our

TABLE I

EVALUATION ON MM-SAFETYBENCH: WE REPORT THE ASR↓ ACROSS SEVEN PROHIBITED CATEGORIES, USING LLAMAGUARD-7B AS THE JAILBREAK CLASSIFIER. BOLD INDICATES THE LOWEST ASR PER COLUMN. SMOOTHGUARD USES RANDOMIZED SMOOTHING ($\sigma = 0.1$; 9 NOISY + 1 ORIGINAL, MAJORITY VOTE).

Model	Method	Illegal Activity	Hate Speech	Malware Generation	Physical Harm	Economic Harm	Fraud	Sex	Avg↓
Qwen2.5-VL-7B	Original	0.0412	**0.0000**	0.0909	0.0417	**0.0000**	0.0130	0.0550	0.0345
	SmoothGuard	**0.0000**	**0.0000**	**0.0682**	**0.0139**	**0.0000**	**0.0000**	**0.0459**	**0.0183**
LLaVA-1.5-7B	Original	0.5876	0.2945	0.3409	0.5208	0.0984	0.5130	0.5413	0.4138
	SmoothGuard	**0.2887**	**0.1411**	**0.2727**	**0.2986**	**0.0164**	**0.1688**	**0.4587**	**0.2350**

experiments build upon the adversarial image attack strategy proposed by [14], but we observed that the transferability of the perturbed adversarial image is limited. To address this, we modify the original pipeline and generalize the implementation such that any MLLM hosted on the HuggingFace platform can be paired with a corresponding adversarial image by simply specifying the model name.

Protocol and Metric. We adopt randomized noise injection with additive Gaussian noise and set the default noise level to $\sigma = 0.1$. At inference, we perform 10 stochastic evaluations per item and aggregate predictions via majority vote based on our method. To determine whether the unsafe prompt still successfully jailbreaks the model, we use an external safety classifier, LlamaGuard-7B. [30] A sample is counted as an attack success if LlamaGuard flags the model's response as violating the safety policy for the given unsafe prompt. We focus on the ASR and report per-category values.

Results. Table I summarizes ASR across seven prohibited categories on MM-SafetyBench. For Qwen2.5-VL-7B, SmoothGuard consistently lowers ASR relative to the original model, nearly eliminating successful attacks in categories such as Illegal Activity and Physical Harm. Similar improvements are observed on LLaVA-1.5-7B, where average ASR drops substantially across all categories. Beyond the raw reductions, these results highlight that a lightweight defense can generalize across architectures with very different baseline vulnerabilities. Taken together, the findings confirm SmoothGuard's effectiveness as a universal defense for multimodal safety.

C. Utility Analysis

For utility analysis, we conducted experiments in two settings: multiple-choice question answering and sentence-based question answering. Evaluations were performed on both the Qwen2.5-VL-7B and LLaVA-1.5-7B models, comparing performance under Gaussian noise ($\sigma = 0.1$) against the no-noise baseline.

Multiple-choice Question. For evaluating multiple-choice performance, we employ the POPE benchmark, a multimodal dataset comprising 9,000 vision–language pairs. POPE is organized into three categories—*Adversarial*, *Popular*, and *Random*—which are specifically designed to assess hallucination tendencies in vision–language models. These categories provide a rigorous framework for revealing the robustness and reliability of model predictions. To ensure a comprehensive evaluation, we report standard classification metrics, including Accuracy, Precision, Recall, and F1 score.

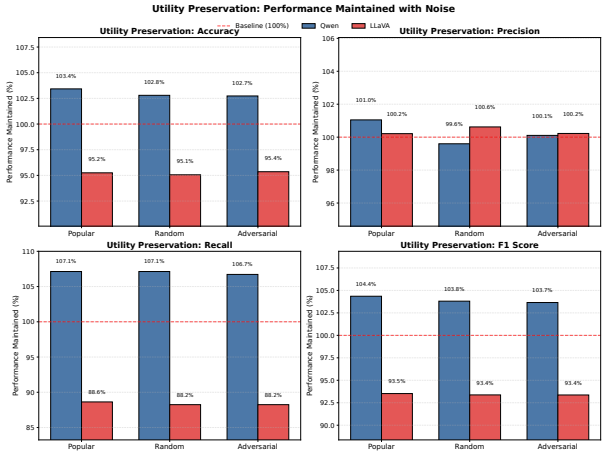

Fig. 3. Utility preservation of Qwen and LLaVA with randomized smoothing noise across different categories. Performance is normalized to the baseline (100%), showing that Qwen improves utility while LLaVA maintains competitive results.

As shown in Fig. 3, our results demonstrate that the introduction of randomized smoothing preserves the utility of both models across different categories. For Qwen, smoothing leads to consistent improvements, with accuracy, precision, recall, and F1 score all exceeding the baseline (100%) across the *Popular*, *Random*, and *Adversarial* settings. In contrast, LLaVA exhibits only a modest decline, but performance remains close to the baseline, confirming that utility is largely maintained even with noise injection. These results indicate that randomized smoothing enhances robustness while ensuring that model utility is not substantially compromised.

Overall, this analysis highlights the practicality of our approach: randomized smoothing provides a reliable defense mechanism that strengthens robustness without sacrificing the effectiveness of MLLMs in utility-driven tasks.

Sentence Question Answering. For sentence-level question answering, we adopt the *LLaVA-Bench (In-the-Wild)*, which consists of diverse image–text pairs divided into three categories: *Conv* (short conversational queries), *Detail* (fine-grained descriptions of the image), and *Complex* (multi-step reasoning and more challenging scenarios). We evaluate both models' answers alongside a reference answer produced by

GPT. To assess performance, we prompt the latest LLM model to act as a judge, rating the quality of model answers relative to the GPT baseline.

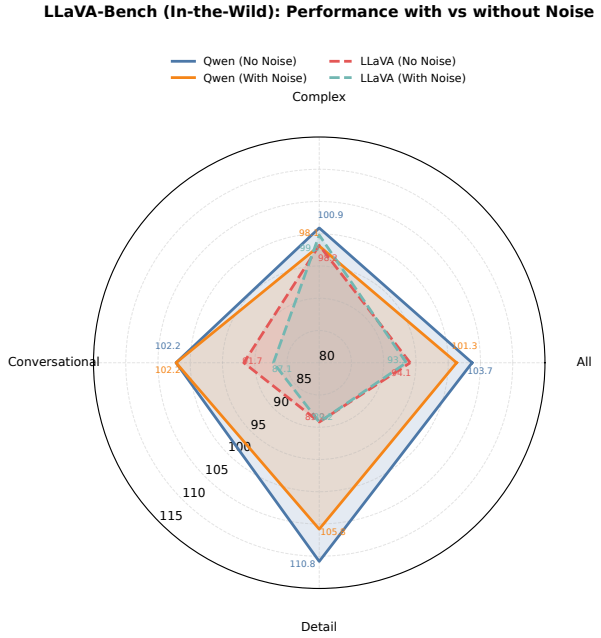

Fig. 4. LLaVA-Bench (In-the-Wild) relative scores of Qwen and LLaVA with and without randomized smoothing noise, compared against GPT-4 (higher is better).

As shown in Fig. 4, introducing randomized smoothing leads to only marginal performance differences for both models. For Qwen, the scores with noise remain close to those without noise across all categories, indicating that utility is effectively preserved. For LLaVA, utility is also largely maintained: there is a small drop in the *Conversational* category, a slight improvement in the *Complex* category, and negligible changes in the *Detail* and *All* categories. Overall, these results suggest that randomized smoothing provides robustness benefits while preserving utility, with only limited trade-offs across different sentence-level tasks.

D. Ablation Study

Before conducting the main experiments, we performed an ablation study to investigate the effect of varying noise levels on model robustness and utility.

Utility For utility, we ran the Qwen2.5-VL-7B model on a subset of the POPE dataset under the adversarial setting and evaluated its accuracy across different Gaussian noise levels.

As illustrated in Fig. 5, model performance remains relatively stable across the tested range, but exhibits a clear optimal region around noise levels of 0.1–0.2. Within this interval, accuracy reaches its peak and demonstrates less fluctuation compared to higher noise values. Beyond this range, particularly after 0.3, we observe a gradual decline in accuracy, suggesting that excessive noise begins to distort the input distribution and impair utility.

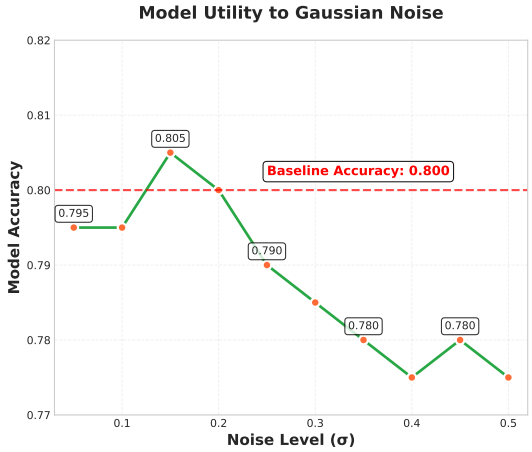

Fig. 5. Model Utility to Gaussian noise on the adversarial setting.

This trend highlights a critical trade-off: introducing a small degree of noise can improve robustness without significantly sacrificing accuracy, but overly large perturbations could possibly degrade model performance. Based on these observations, we select 0.1–0.2 as the optimal range of noise levels for the subsequent experiments.

Robustness To assess sensitivity to the noise magnitude, we conduct an ablation on a representative subset of MM-SafetyBench scenarios while keeping the repeated-evaluation protocol fixed. We vary the noise level from 0.05 to 0.50 in increments of 0.05 and report the ASR as the sole metric.

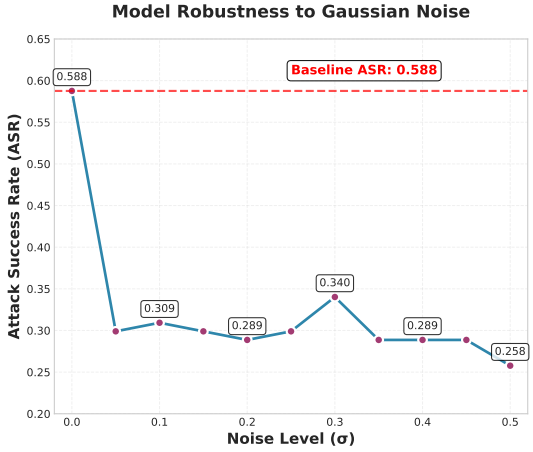

Fig. 6. Model Robustness to Gaussian noise on the adversarial setting.

Fig. 6 shows the change of the ASR. Compared to the baseline ASR of 0.588, even small perturbations substantially improve robustness: ASR drops to 0.309 at $\sigma = 0.10$ and reaches its lowest value of 0.258 at $\sigma = 0.30$. Beyond this point, the curve flattens, with no consistent gains at higher noise levels. These results suggest that moderate noise in the range of 0.10–0.30 achieves the most favorable trade-off, effectively reducing adversarial vulnerability while avoiding unnecessary degradation. The 10-pass majority vote further

stabilizes predictions, reducing variance across runs.

V. Conclusion

In this work, we introduced defense framework for MLLMs by introducing random noise and semantic clustering techniques, targeting adversarial manipulations in continuous modalities such as images and audio. By perturbing inputs with Gaussian noise and aggregating predictions, our method effectively enhances robustness against adversarial attacks while preserving task utility. Experiments on POPE, Bench-in-the-Wild, and MM-SafetyBench demonstrate that randomized smoothing mitigates hallucinations and adversarial vulnerabilities with only marginal performance degradation under benign conditions, achieving results that remain competitive with strong baselines. Ablation studies further reveal that modest noise levels (0.1–0.2) strike the best balance between robustness and utility. These findings highlight randomized smoothing as a lightweight, model-agnostic, and practical defense for securing MLLMs. In future work, we will extend our experiments to the audio modality, further validating the effectiveness of randomized smoothing beyond vision–language tasks.

References

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

[2] D. Chen, R. Chen, S. Zhang, Y. Liu, Y. Wang, H. Zhou, Q. Zhang, Y. Wan, P. Zhou, and L. Sun, "Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark," *arXiv preprint arXiv:2402.04788*, 2024.

[3] X. Zhang, H. Wen, J. Wu, P. Qin, H. Xue', and L. Nie, "Differential-perceptive and retrieval-augmented mllm for change captioning," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 4148–4157.

[4] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[5] Q. Ye, H. Xu, J. Ye, M. Yan, A. Hu, H. Liu, Q. Qian, J. Zhang, and F. Huang, "mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2024, pp. 13 040–13 051.

[6] T. Gao, P. Chen, M. Zhang, C. Fu, Y. Shen, Y. Zhang, S. Zhang, X. Zheng, X. Sun, L. Cao *et al.*, "Cantor: Inspiring multimodal chain-of-thought of mllm," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 9096–9105.

[7] X. Liu, Y. Zhu, Y. Lan, and Y. Yang, Chao andF Qiao, "Safety of multimodal large language models on images and text," *arXiv preprint arXiv:2402.00357*, 2024.

[8] Y. Chen, H. Li, Z. Zheng, and Y. Song, "Bathe: Defense against the jailbreak attack in multimodal large language models by treating harmful instruction as backdoor trigger," *arXiv preprint arXiv:2408.09093*, 2024.

[9] X. Du, R. Ghosh, R. Sim, A. Salem, V. Carvalho, E. Lawton, Y. Li, and J. W. Stokes, "Vlmguard: Defending vlms against malicious prompts via unlabeled data," *arXiv preprint arXiv:2410.00296*, 2024.

[10] Z. Zhou, Q. Wang, M. Jin, J. Yao, J. Ye, W. Liu, W. Wang, X. Huang, and K. Huang, "Mathattack: Attacking large language models towards math solving ability," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 19 750–19 758.

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[12] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[13] S. Wang, Z. Long, Z. Fan, and Z. Wei, "From llms to mllms: Exploring the landscape of multimodal jailbreaking," *arXiv preprint arXiv:2406.14859*, 2024.

[14] X. Qi, K. Huang, A. Panda, P. Henderson, M. Wang, and P. Mittal, "Visual adversarial examples jailbreak aligned large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 19, 2024, pp. 21 527–21 536.

[15] J. Wen, X. Wu, S. Zhao, Y. Jia, and Y. Li, "Investigating vulnerabilities and defenses against audio-visual attacks: A comprehensive survey emphasizing multimodal models," *arXiv preprint arXiv:2506.11521*, 2025.

[16] Y. Zhao, T. Pang, C. Du, X. Yang, C. Li, N.-M. M. Cheung, and M. Lin, "On evaluating adversarial robustness of large vision-language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[17] G. Chen, F. Song, Z. Zhao, X. Jia, Y. Liu, Y. Qiao, and W. Zhang, "Audiojailbreak: Jailbreak attacks against end-to-end large audio-language models," *arXiv preprint arXiv:2505.14103*, 2025.

[18] B. Ma, H. Guo, Z. J. Luo, and R. Duan, "Audio jailbreak attacks: Exposing vulnerabilities in speechgpt in a white-box framework," *arXiv preprint arXiv:2505.18864*, 2025.

[19] Z. Song, Q. Jiang, M. Cui, M. Li, L. Gao, Z. Zhang, Z. Xu, Y. Wang, C. Wang, G. Ouyang *et al.*, "Audio jailbreak: An open comprehensive benchmark for jailbreaking large audio-language models," *arXiv preprint arXiv:2505.15406*, 2025.

[20] Z. Qian, K. Huang, Q.-F. Wang, and X.-Y. Zhang, "A survey of robust adversarial training in pattern recognition: Fundamental, theory, and methodologies," *Pattern Recognition*, vol. 131, p. 108889, 2022.

[21] C. Lyu, K. Huang, and H.-N. Liang, "A unified gradient regularization family for adversarial examples," in *2015 IEEE international conference on data mining*. IEEE, 2015, pp. 301–309.

[22] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter, "Certified adversarial robustness via randomized smoothing," *arXiv preprint arXiv:1902.02918*, 2019.

[23] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," *arXiv preprint arXiv:1705.07204*, 2017.

[24] R. Peri, S. M. Jayanthi, S. Ronanki, A. Bhatia, K. Mundnich, S. Dingliwal, N. Das, Z. Hou, G. Huybrechts, S. Vishnubhotla *et al.*, "Speechguard: Exploring the adversarial robustness of multimodal large language models," *arXiv preprint arXiv:2405.08317*, 2024.

[25] A. Djanibekov, N. Mukhituly, K. Inui, H. Aldarmaki, and N. Lukas, "Spirit: Patching speech language models against jailbreak attacks," *arXiv preprint arXiv:2505.13541*, 2025.

[26] J. Sun, C. Wang, J. Wang, Y. Zhang, and C. Xiao, "Safeguarding vision-language models against patched visual prompt injectors," *arXiv preprint arXiv:2405.10529*, 2024.

[27] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[28] Y. Bai, X. Yan, Y. Jiang, S.-T. Xia, and Y. Wang, "Clustering effect of adversarial robust models," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 590–29 601, 2021.

[29] X. Liu, Y. Zhu, J. Gu, Y. Lan, C. Yang, and Y. Qiao, "Mm-safetybench: A benchmark for safety evaluation of multimodal large language models," in *European Conference on Computer Vision*. Springer, 2024, pp. 386–403.

[30] H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, D. Testuggine *et al.*, "Llama guard: Llm-based input-output safeguard for human-ai conversations," *arXiv preprint arXiv:2312.06674*, 2023.