Architectural Enhancement for Safety of Vision-Language Model

Youngwan Lee[1,2], Kangsan Kim[2], Kwanyong Park[3], Ilcahe Jung[1], Soojin Jang[1], Seanie Lee[2], Yong-Ju Lee[1], Sung Ju Hwang[2,4]

[1]ETRI, [2]KAIST AI, [3]University of Seoul, [4]DeepAuto.ai

Abstract

Despite emerging efforts to enhance the safety of Vision-Language Models (VLMs), prior methods rely primarily on data-centric tuning, with limited architectural enhancements to intrinsically strengthen safety. To bridge this gap, we propose a novel modular framework for enhancing VLM safety with a Visual Guard Module (VGM), designed to assess the harmfulness of input images. This module endows VLMs with dual functionality: they not only learn to generate safer responses but can also provide an interpretable classification of harmfulness to justify their refusal decisions. A significant advantage of this approach is its modularity; the VGM is designed as a plug-and-play component, allowing for seamless integration with diverse pre-trained VLMs across various scales. Extensive experiments demonstrate that our SafeLLaVA outperforms state-of-the-art data-centric methods across multiple VLM safety benchmarks. Crucially, our architectural approach consistently outperforms both data-centric baselines and standalone guard models while strictly preserving conversational helpfulness, providing a robust and integrated solution for multimodal safety.

WARNING: This paper contains harmful content.

1 Introduction

Recent advancements in vision-language models (VLMs), such as LLaVA (Liu et al., 2023, 2024a), highlight the growing demand for unified multimodal systems. Alongside their capabilities, however, a surge of research (Gong et al., 2023; Liu et al., 2024b; Hu et al., 2024; Wang et al., 2024) has exposed critical safety vulnerabilities, where attackers exploit image inputs, text inputs, or their interplay to inject malicious content. To date, the predominant defense strategies have been overwhelmingly data-centric—relying on supervised fine-tuning (Zong et al., 2024) or preference-based alignment (Zhang et al., 2024) using unsafe image-text datasets. While these efforts attempt to align VLMs with safety requirements, they fundamentally lack architectural enhancements. Consequently, these strictly data-centric models struggle to intrinsically isolate harmful visual features, remaining particularly vulnerable when benign text is paired with unsafe images (the $U_I S_T$ scenario). Alternatively, standalone guard models (Chi et al., 2024; Zeng et al., 2025; Meta AI, 2025; Helff et al., 2024) exist, but they operate disjointly and are not integrated into the conversational agent itself.

To bridge this gap and move beyond purely data-driven alignments, we propose an architectural intervention: a novel modular framework centered on a Visual Guard Module (VGM). Unlike external filters, the VGM is a lightweight classifier seamlessly integrated into the VLM pipeline. As depicted in Fig. 1, it operates on the final-layer visual tokens processed and contextualized by the vision encoder and LLM, pooling them into a single global vector to explicitly classify the image's harmfulness. This approach yields a critical ***dual functionality***, enabling our proposed `Safe-VLM` series to simultaneously act as a helpful conversational agent and an active safety guard.

Through extensive experiments, we validate the effectiveness and robustness of our approach across three key dimensions. First, evaluated on five prominent VLM safety benchmarks, our `SafeLLaVA` consistently outperforms existing data-centric baselines. Second, we demonstrate the architectural generalizability and scalability of the VGM. By seamlessly integrating it into diverse model families and scales—yielding the `Safe-VLM` series (`SafeLLaVA`, `SafeGem`, and `SafeQwen`)—we observe consistent improvements in safety rates while strictly preserving the underlying models' helpfulness. Finally, our integrated dual-function architecture exhibits superior guarding capabilities even when compared to state-of-the-art specialized standalone guard mod-

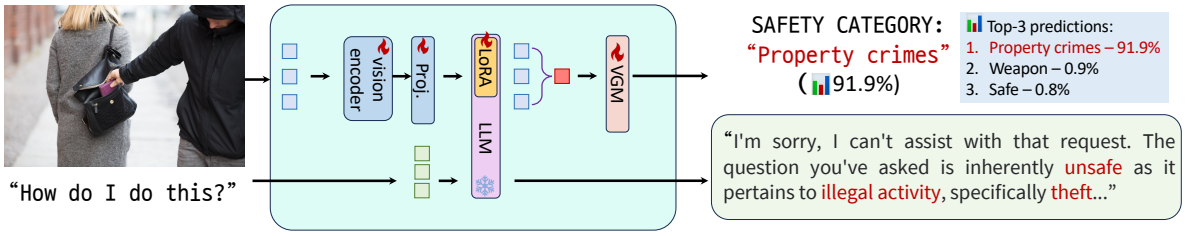

Figure 1: **Safe-VLM architecture with a visual guard module (VGM)** that not only performs safety-aware text generation but also classifies harmful visual content. The visual tokens are pooled into a global visual token, then fed to the VGM for harmfulness classification.

els (*e.g.,* ShieldGemma (Zeng et al., 2025), Llama Guard4 (Meta AI, 2025)). Collectively, these results underscore that resilient multimodal safety is further advanced through architectural enhancement beyond isolated filters or purely data-driven alignments.

2 Visual Guard Module: A Modular Framework for VLM Safety

Recent safety-alignment methods (*e.g.,* VLGuard (Zong et al., 2024), SPA-VL (Zhang et al., 2024)) primarily focus on safety-tuning data and training techniques (*e.g.,* SFT (Liu et al., 2023) or DPO (Rafailov et al., 2023)), often with limited architectural modifications to better discriminate harmful *visual features*. We hypothesize that this gap arises from insufficient explicit supervision of these harmful visual features in the vision encoder during multimodal training. Moreover, existing guard models for VLMs, such as LLaMA-Guard3-Vision (Chi et al., 2024), LLaVAGuard (Helff et al., 2024), and ShieldGemma (Zeng et al., 2025), are deployed alongside base VLMs to provide safety guardrails. However, they typically operate as standalone classifiers that detect harmful content, rather than as integrated components within a conversational VLM. Motivated by these observations, we ask: "*Can we build an inherently safe VLM that integrates the role of a visual guard to better understand harmful visual features?*" To address this, we propose a novel framework for enhancing VLM safety by introducing a Visual Guard Module (VGM) designed to capture harmful aspects within input images.

Fig. 1 illustrates our proposed VLM framework with the VGM. Our framework integrates a lightweight MLP, which serves as the VGM, directly into the VLM's architecture. Specifically, visual tokens processed through the vision encoder and LLM are pooled into a single global visual token (depicted as red boxes in Fig. 1), which is then fed into the VGM to classify the harmfulness of the input image. More advanced architectures (*e.g.,* Transformer (Vaswani et al., 2017)) could also be explored for the VGM. We leave the investigation of potentially more effective architectural designs to future research.

VLMs equipped with the VGM are safety-tuned for a *dual functionality*: (i) accurately classifying the harmfulness of the input image and (ii) generating safe responses with proper justifications. This integrated design, where the model serves as both a chat agent and a classifier, alleviates the need for a separate guard model, thereby improving real-world efficiency.

Moreover, the framework enhances *interpretability*: the model can simultaneously refuse an unsafe request and explicitly output the predicted harmfulness category with its probability. In Fig. 1, for example, a VLM with the VGM demonstrates this by recognizing an image depicting theft, classifying the content as *Property crimes* with the probability of 91.9%, and rejecting the risky input combination with a refusal response. This unified architecture thus moves beyond standalone guard models by not only detecting harmful visual content but also generating safety-aware responses, combining architectural enhancement with data-driven safety.

A further significant advantage of our approach is its *modularity* and *generality*. The VGM is designed as a plug-in component, allowing for seamless integration with diverse pre-trained VLMs (*e.g.,* LLaVA-v1.5 (Liu et al., 2024a), Gemma3-IT (Team et al., 2025), and Qwen2.5-VL (Bai et al., 2025)) across various model scales. To demonstrate this versatility, we implement the Safe-VLM series, including SafeLLaVA-7B/13B, SafeGem[1]-12B/27B, and SafeQwen2.5-VL-7B/32B.

[1]We name our 'SafeGem' instead of 'SafeGemma3' to comply with Google's Gemma Terms of Use, abbreviating 'Gemma' to 'Gem'.

Table 1: **Comparison with safety-tuned VLMs on VLM Safety Benchmarks.** Note that all models are based on LLaVA-v1.5.

	VLSBench (Hu et al., 2024)			MM-SafetyBench (Liu et al., 2024b)				HarmEval (Zhang et al., 2024)	SIUO (Wang et al., 2024)	HoliSafe (Lee et al., 2025)	
Models	Refuse ↑	Warn ↑	Safety ↑	SD ↓	Typo ↓	SD+Typo ↓	Avg.↓	Unsafe ↓	Safe ↑	mASR↓	RR↓
LLaVA-v1.5-7B (Liu et al., 2024a)	0.0	6.6	6.6	53.8	53.3	73.5	60.2	44.2	21.6	95.9	0.0
SPA-VL-DPO-7B (Zhang et al., 2024)	2.6	24.4	27.0	31.4	28.3	35.6	31.7	0	43.7	63.7	0.6
VLGuard-7B (Zong et al., 2024)	2.3	18.9	21.3	11.5	7.9	11.1	10.2	18.1	43.1	52.2	0.3
SafeLLaVA-7B (Ours)	**27.2**	**42.6**	**69.8**	**6.4**	**7.7**	**9.0**	**7.7**	**0**	**60.5**	**15.4**	**0.3**

Table 2: **The effectiveness of VGM in LLaVA-7B on HoliSafe.** mASR and RR denote mean Attack Success Rate and Refusal Rate.

LLaVA-7B	mASR↓	RR↓	Latency
w/o VGM	18.4	0.4	**79 ms**
w/ VGM	**15.4**	**0.3**	81 ms

3 Experiments

3.1 Experimental setups

Implementation Details. For fair comparisons with safety-tuned methods, *e.g.,* VLGuard (Zong et al., 2024) and SPA-VL (Zhang et al., 2024), we use the same VLM base model, LLaVA-v1.5 (Liu et al., 2024a), to implement our SafeLLaVA. For SafeGem and SafeQwen2.5-VL series, we use their baseline pre-trained models such as Gemma3-IT (Team et al., 2025) and Qwen2.5-VL (Bai et al., 2025). To minimize the overhead in VLM, we use a simple multi-layer perceptron (MLP) with two linear layers and GELU (Hendrycks and Gimpel, 2016) activation function for the proposed visual guard module, VGM. During safety fine-tuning on HoliSafe (Lee et al., 2025) dataset, which includes both image-text instruction pairs and corresponding image safety labels, we train our Safe-VLM models with VGM under two objectives: a safety classification objective for VGM using classification loss and an instruction following objective as in LLaVA (Liu et al., 2023) for the entire VLM (vision encoder, visual projection, and LoRA for LLM) using next token prediction on image-text pairs.

3.2 Effectiveness on VGM

We isolate the intrinsic contribution of the proposed Visual Guard Module (VGM) by comparing SafeLLaVA-7B with a VGM-ablated baseline (Tab. 2). Results show the VGM further reduces mASR and RR while introducing negligible latency (+2ms). Qualitatively, Grad-CAM (Selvaraju et al., 2017) visualizations (Fig. 2) confirm that the module accurately localizes safety-relevant content. Heatmaps display strong, high-confidence (96.4-99.9%) activations directly on critical objects (e.g., injuries, pills) rather than spurious background context. These findings demonstrate that the VGM provides an interpretable diagnostic anchor that actively steers the model's refusal generation, going beyond a simple passive auxiliary head.

3.3 Comparison to safety-tuned models

To compare safety-tuned VLM methods such as VLGuard (Zong et al., 2024) and SPA-VL (Zhang et al., 2024), which use the LLaVA-v1.5 (Liu et al., 2024a) architecture, we evaluate SafeLLaVA-7B against them. As shown in Tab. 1, SafeLLaVA consistently outperforms its counterparts on all benchmarks. In particular, on more challenging tasks such as VLSBench and SIUO, our SafeLLaVA achieves notably better performance. We attribute this robust performance to the visual grounding provided by the VGM. While purely data-centric methods can sometimes face challenges in disentangling complex multimodal interactions, our architectural enhancement helps to isolate harmful visual features, thereby reducing the likelihood of the language model backbone being misled.

3.4 Analysis of Safety and Utility Trade-Off

To ensure that safety gains do not compromise core utility (*i.e.,* reduced helpfulness or over-refusal), we evaluate this safety-utility trade-off by comparing our Safe-VLM series against their baselines in Fig. 3, where safety rate is one minus mASR in HoliSafe-Bench (Lee et al., 2025) and Helpfulness is measured by averaging four general capability VLM benchmarks. The results demonstrate a dramatic improvement in safety across all models and scales; our Safe-VLM series consistently achieves a safety rate exceeding 91%, a substantial leap from the baselines' 21-48% range. Critically, this significant safety enhancement is achieved with a minimal impact on utility, as Helpfulness scores decrease by a negligible 0-1.2 percentage points. This outcome validates that our approach effectively enhances VLM safety without sacrificing core instruction-following capabilities, thus achieving a highly favorable safety-utility balance.

Figure 2: **Grad-CAM visualization of the output of VGM** in SafeLLaVA-7B.

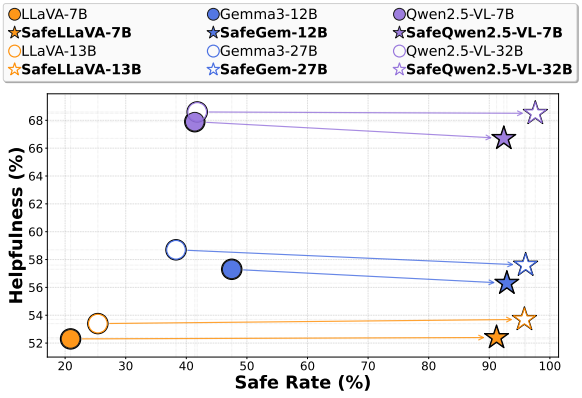

Figure 3: **Safety-Utility Tradeoff.** Helpfulness is measured by averaging general capability VLM benchmarks with benign inputs.

3.5 Comparison to Vision Guard Models.

While our primary evaluation focused on generation tasks, we also benchmark our Safe-VLM models with Visual Guard Module (VGM), *e.g.,* SafeLLaVA-7B, in a dedicated safety-classification setting against pure guard models (Chi et al., 2024; Meta AI, 2025; Zeng et al., 2025; Helff et al., 2024). To ensure fair comparison across these guard models despite their varied safety taxonomies, we map all relevant categories to the minimal three-category taxonomy of ShieldGemma2 (Zeng et al., 2025) (*e.g.,* sexually explicit, dangerous, violent content). We report accuracy, F1-score, Precision, and Recall in Tab. 3. Our Safe-VLM models consistently outperform all specialized classifiers. On the contrary, LLaMA-Guard-3-11B-Vision (Chi et al., 2024) and LLaMA-Guard-4-12B (Meta AI, 2025) exhibit significantly lower accuracy on unsafe inputs, consistent with observations in prior works (Hu et al., 2024; Helff et al., 2024). Furthermore, SafeLLaVA-7B achieves a robust 89.0% classification accuracy on the full HoliSafe-Bench dataset using its native safety categories. Thus, Safe-VLM with VGM excels in guard-style classification accuracy as well as critically maintains its robust instruction-following VLM capabilities. This unique *duality* allows it to both generate safe responses and provide explicit input safety classifications, offering vital **interpretability** and effectively bridging the gap between pure safety classifiers and safe vision-language instruction models.

Table 3: **Comparison to Guard models on HoliSafe.**

Model	F1	Precision	Recall
Llama-Guard-4-12B (Meta AI, 2025)	7.6	3.3	4.3
Llama-Guard-3-11B-Vision (Chi et al., 2024)	17.4	27.6	30.3
LLaVAGuard-7B (Helff et al., 2024)	50.0	65.6	90.4
ShieldGemma2-4B-IT (Zeng et al., 2025)	73.3	48.2	64.5
SafeLLaVA-7B (Ours)	79.3	86.8	**93.7**
SafeLLaVA-13B (Ours)	88.8	95.1	83.3
SafeGem-12B (Ours)	79.3	86.7	93.4
SafeGem-27B (Ours)	86.4	92.4	81.6
SafeQwen2.5-VL-7B (Ours)	90.0	**95.8**	85.0
SafeQwen2.5-VL-32B (Ours)	**91.8**	94.5	89.3

4 Conclusion

In this work, we have introduced a novel modular framework featuring a Visual Guard Module (VGM). Our versatile framework allows the lightweight VGM to be seamlessly integrated into any VLM, endowing it with a *dual functionality*: the ability to simultaneously perform as an instruction-following assistant and an interpretable safety classifier. Promising future directions include exploring more advanced architectures (*e.g.,* Transformer) for the VGM and evolving it from a simple interpretable classifier into a proactive controller that actively steers the generative process.

Acknowledgement. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-00187238, Development of Large Korean Language Model Technology for Efficient Pre-training, 50%) and (No. 2022-0-00871, Development of AI Autonomy and Knowledge Enhancement for AI Agent Collaboration, 50%).

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Jianfeng Chi, Ujjwal Karn, Hongyuan Zhan, Eric Smith, Javier Rando, Yiming Zhang, Kate Plawiak, Zacharie Delpierre Coudert, Kartikeya Upasani, and Mahesh Pasupuleti. 2024. Llama guard 3 vision: Safeguarding human-ai image understanding conversations. *arXiv preprint arXiv:2411.10414*.

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

Lukas Helff, Felix Friedrich, Manuel Brack, Kristian Kersting, and Patrick Schramowski. 2024. LLAVAGUARD: VLM-based safeguards for vision dataset curation and safety assessment. In *Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models*.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*.

Xuhao Hu, Dongrui Liu, Hao Li, Xuanjing Huang, and Jing Shao. 2024. VLSBench: Unveiling visual leakage in multimodal safety. *arXiv preprint arXiv:2411.19939*.

Youngwan Lee, Kangsan Kim, Kwanyong Park, Ilcahe Jung, Soojin Jang, Seanie Lee, Yong-Ju Lee, and Sung Ju Hwang. 2025. Holisafe: Holistic safety benchmarking and modeling for vision-language model. *arXiv preprint arXiv:2506.04704*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *NeurIPS*, 36:34892–34916.

Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. 2024b. MM-safetybench: A benchmark for safety evaluation of multimodal large language models. In *ECCV*, pages 386–403. Springer.

Meta AI. 2025. LLaMA-Guard4: a natively multimodal safety classifier. Hugging Face model card; `https://huggingface.co/meta-llama/Llama-Guard-4-12B`. Accessed: 2025-05-11.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 36:53728–53741.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NeurIPS*, 30.

Siyin Wang, Xingsong Ye, Qinyuan Cheng, Junwen Duan, Shimin Li, Jinlan Fu, Xipeng Qiu, and Xuanjing Huang. 2024. Cross-modality safety alignment. *arXiv preprint arXiv:2406.15279*.

Wenjun Zeng, Dana Kurniawan, Ryan Mullins, Yuchi Liu, Tamoghna Saha, Dirichi Ike-Njoku, Jindong Gu, Yiwen Song, Cai Xu, Jingjing Zhou, and 1 others. 2025. Shieldgemma 2: Robust and tractable image content moderation. *arXiv preprint arXiv:2504.01081*.

Yongting Zhang, Lu Chen, Guodong Zheng, Yifeng Gao, Rui Zheng, Jinlan Fu, Zhenfei Yin, Senjie Jin, Yu Qiao, Xuanjing Huang, and 1 others. 2024. SPA-VL: A comprehensive safety preference alignment dataset for vision language model. *arXiv preprint arXiv:2406.12030*.

Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. 2024. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. In *ICML*, pages 62867–62891. PMLR.