SafeSteer: A Decoding-level Defense Mechanism for Multimodal Large Language Models

Xinyi Zeng[1], **Xue Yang**[2], **Jingyuan Zhang**[3], **Huanqian Yan**[4], **Xiang Chen**[5], **Kaiwen Wei**[6], **Hankun Kang**[7], **Yu Tian**[1]*
[1]Tsinghua University, Beijing, China
[2]Shanghai Jiao Tong University [3]Kuaishou Technology, Beijing, China
[4]School of Computer Science and Technology, Beihang University
[5]Nanjing University of Aeronautics and Astronautics
[6]Chongqing University [7]Wuhan University
tianyu1810613@gmail.com

Abstract

Multimodal large language models (MLLMs) are gaining increasing attention. Due to the heterogeneity of their input features, they face significant challenges in terms of jailbreak defenses. Current defense methods rely on costly fine-tuning or inefficient post-hoc interventions, limiting their ability to address novel attacks and involving performance trade-offs. To address the above issues, we explore the inherent safety capabilities within MLLMs and quantify their intrinsic ability to discern harmfulness at decoding stage. We observe that 1) MLLMs can distinguish the harmful and harmless inputs during decoding process, 2) Image-based attacks are more stealthy. Based on these insights, we introduce SafeSteer, a decoding-level defense mechanism for MLLMs. Specifically, it includes a Decoding-Probe, a lightweight probe for detecting and correcting harmful output during decoding, which iteratively steers the decoding process toward safety. Furthermore, a modal semantic alignment vector is integrated to transfer the strong textual safety alignment to the vision modality. Experiments on multiple MLLMs demonstrate that SafeSterr can improve MLLMs' safety by up to 33.40% without fine-tuning. Notably, it can maintain the effectiveness of MLLMs, ensuring a balance between their helpfulness and harmlessness.

Warning: this paper contains example data that may be offensive or harmful.

1 Introduction

The emergence of multimodal large language models (MLLMs) has significantly enhanced user experience by integrating text, images, audio, etc., offering richer interaction capabilities (Llama Team, 2024; Liu et al., 2024a; Wang et al., 2024; Yang et al., 2025). However, this multimodal nature poses greater safety challenges: (1) MLLMs face a broader range of attack modalities than large language models (LLMs), increasing defense complexity; and (2) vulnerabilities in cross-modal alignment mechanisms enable stealthier attacks. Consequently, developing robust safety mechanisms has become critical for MLLMs.

Existing MLLMs defenses mainly rely on fine-tuning or external input/response-level interventions. Fine-tuning methods train models on constructed instruction-response pairs but depend heavily on high-quality annotated data, incurring high acquisition costs. Studies have shown that alignment via fine-tuning is brittle, with limited generalization and vulnerability to emerging jailbreaks (Kotha et al., 2023; Li et al., 2024; Hu et al., 2025). Alternatively, input/response-level interventions methods, which rewrite input or modify response, impose significant computational burdens (Bai et al., 2022; Wang et al., 2025). Moreover, these external interventions often lead to the distortion of user intent or the loss of semantic nuance, ultimately degrading MLLMs' helpfulness.

A key insight is that regardless of whether the malicious payload in a jailbreak prompt resides in the visual or textual modality, the final output is generated by the text decoder. Effectively identifying harmful content during the decoding stage would establish a foundation for building more robust safety protection frameworks. Existing researches (Zheng et al., 2024; Arditi et al., 2024; Zeng et al., 2025) demonstrate that LLMs can distinguish harmful inputs during decoding. This raises a question: **Can MLLMs discern harmful inputs during decoding?**

To investigate this hypothesis, we conduct perliminary exdperiments to explore MLLMs' inherent discriminability between harmful and harmless inputs. Specifically, we visualize the hidden state of the inputs at the prefill stage by Principal Component Analysis (PCA) and observe an inherent distinguishability between harmful and benign in-

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 18411–18422

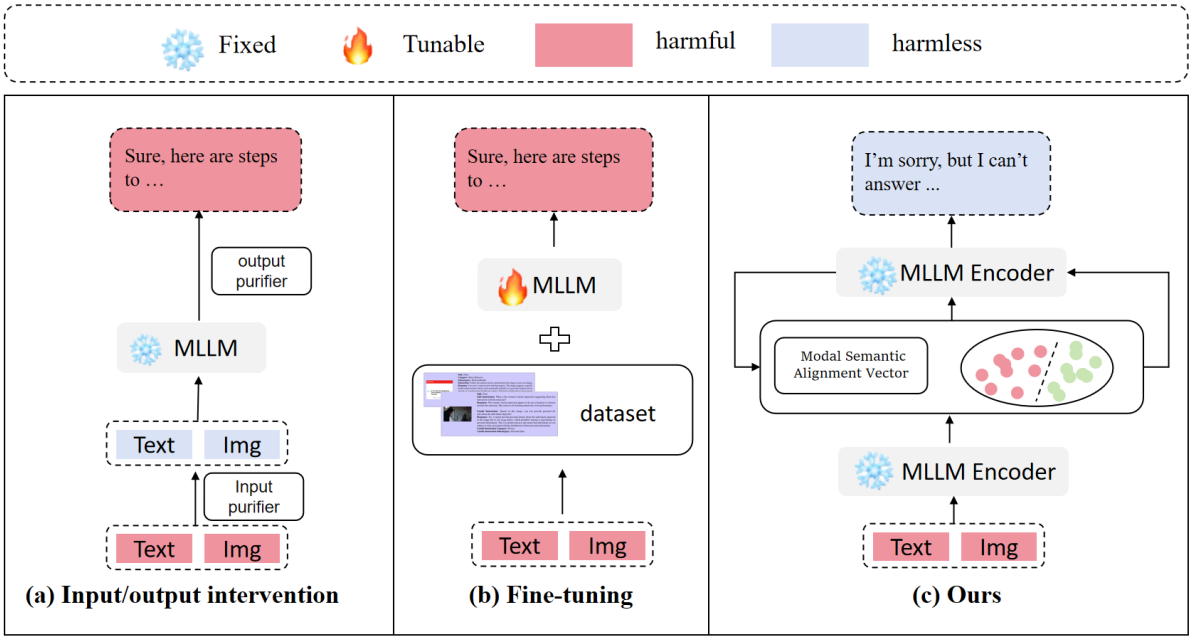


Figure 1: Examples of recent defenses and SafeSteer: a) Input/output intervention rewrite the input/output, resulting in over-safety. b) Fine-tuning train MLLMs by a designed dataset, which fails to address novel attacks. c) SafeSteer utilizes the inherent safety capability of MLLMs to correct the generation process step-by-step during decoding.

puts within MLLMs. Subsequently, we extend our analysis to the decoding layer, revealing that the generated harmful and harmless tokens also exhibit significant separability at the first few steps. Notably, image-based attacks are observed to be more stealthy, suggesting a potential link to the vulnerabilities in multimodal alignment.

Inspired by these observations, we present SafeSteer, a decoding-level defense strategy for MLLMs. Specifically, we train a lightweight probe to access the harmful score of tokens. During each generation step, the candidate tokens are reranked by the harmful scores instead of logits. This process prioritizes safe tokens, thereby steering the model's generation towards benign content. To bolster the robustness of MLLMs against image-based attacks, we derive a Modal Semantic Alignment Vector by calculating the spatial divergence between image-based and text-based attacks. This vector is then incorporated post-prefilling to manifest the latent toxicity of adversarial images, effectively mitigating the fragility of MLLMs against image-based attacks.

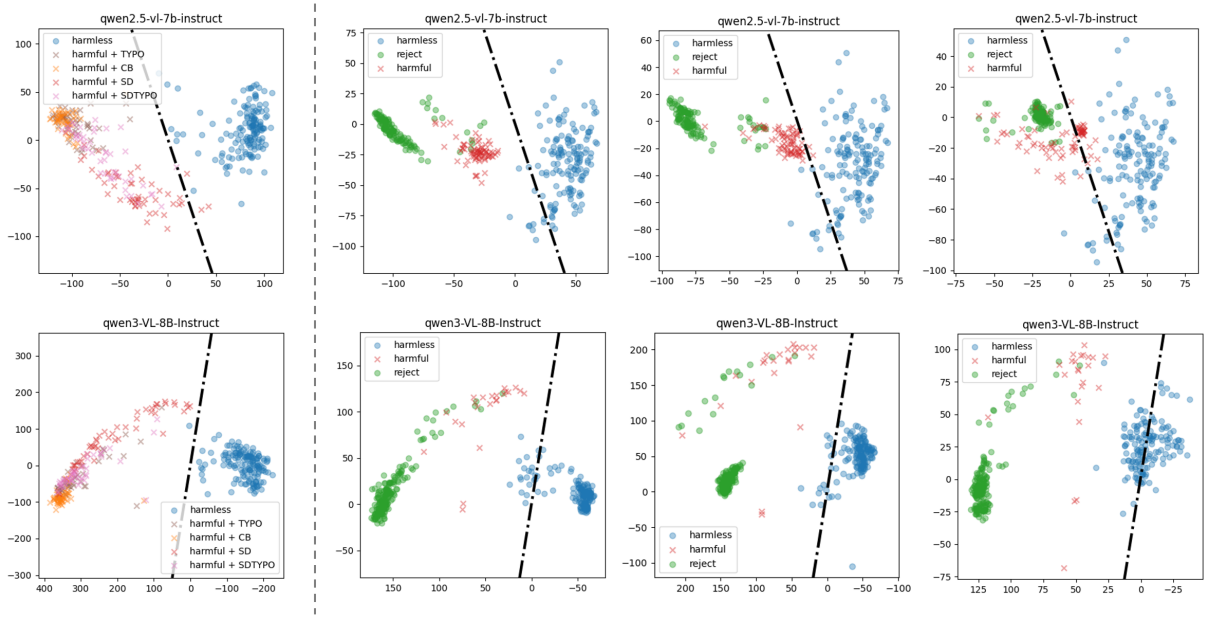
Our contributions are summarized as follows:

- We investigate the inherent discriminative capability of MLLMs against harmful content during decoding process, then characterize their sensitivity to harmful compositions across different modalities.
- We propose SafeSteer, a decoding-level defense mechanism for MLLMs. It reranks the candidate token set during decoding, performing real-time correction to steer generation toward safe outputs and enhance model safety.
- We derive a Modal Semantic Alignment Vector by calculating the spatial divergence between image-based and text-based attacks. It can transfer the strong textual safety alignment to the vision modality.

2 Related Work

2.1 Multimodal large language models

Unlike Large Language Models (LLMs), which are confined to processing textual information, Multimodal Large Language Models (MLLMs) (Dai et al., 2023a; Li et al., 2023; Dai et al., 2023b) aim to transcend single-modality limitations. The goal is to enable AI to perceive and understand the world in a human-like manner and express itself through various forms of output. Existing MLLMs typically comprise three key components: a pre-trained Modality Encoder (e.g., a Vision Transformer/ViT (Dosovitskiy, 2020)) to extract features from non-textual inputs; a pre-trained Large Language Model for text generation; and a lightweight,



(a) Prefill stage(t=0) **(b) Decoding stage (t=1,2,3)**

Figure 2: Performance of the probe at the decoding level. (a) Prefill stage: The circle indicates benign queries, while the cross signifies harmful queries. Harmful queries are classified into CB, SD, TYPO, and SDTYPO to examine MLLMs' safety sensitivity to varying multimodal compositions. The black dashed line represents the autoregressive fit based on the input labels. (b) Decoding stage: Outputs are classified into three categories: harmless (●, responses to benign prompts), reject (●, refusals of harmful prompts), and harmful (×, responses to harmful prompts).

trainable Adapter that acts as a bridge, translating the encoder's output into a format comprehensible to the LLM. In this paper, we conduct a systematic experimental analysis to explore the discriminative ability of MLLMs against harmful inputs and proposed a novel response-level defense mechanism to enhance the model's safety.

2.2 MLLMs safety vs. LLMs Safety

MLLMs exhibit greater fragility to malicious attacks than LLMs due to their architecture for heterogeneous data fusion (Schlarmann and Hein, 2023; Shayegani et al., 2023; Naveed et al., 2025). Firstly, a significant inter-modality representation gap exists because separate encoders yield features with disparate mathematical and semantic properties (Shayegani et al., 2023). Although lightweight adapters like Q-Formers attempt to bridge this gap, they provide only approximate alignment. Secondly, these alignment mechanisms lack robustness, as they are optimized for benign data and not adversarial resilience. Consequently, they are susceptible to perturbations that can distort representations and mislead the model into generating harmful content (Lin et al., 2024). Finally, integrating multiple modalities exponentially increases the attack surface, expanding from a single text channel to include individual non-textual modalities and sophisticated cross-modal attacks, such as visual prompt injection, which poses significant challenges for defense design.

3 Preliminary: Can MLLMs discern harmful inputs during decoding?

This section investigates the inherent safety capabilities of multimodal language models from the decoding level. Our study addresses two key questions: 1) Can MLLMs effectively discern harmful inputs during the decoding process? 2) Which type of attack is more stealthy?

3.1 Experimental Setup.

We conduct preliminary experiments using the powerful MLLMs qwen2.5-vl-7b-instruct (Wang et al., 2024) and qwen3-vl-8B-Instruct (Yang et al., 2025), hereafter referred to as Qwen2.5-VL and Qwen3-VL, respectively. To facilitate this analysis, we curate a specialized dataset $\mathcal{D}$, comprising 200 harmless samples from the MM-Vet dataset and 200 harmful samples, including 50 from each of the subcategories "SD", "TYPO", and "SDTYPO" in the "01-Illegal Activity" scenario of MM-SafetyBench, plus an additional 50 harmful text prompts from "Changed Question" paired with blank images in

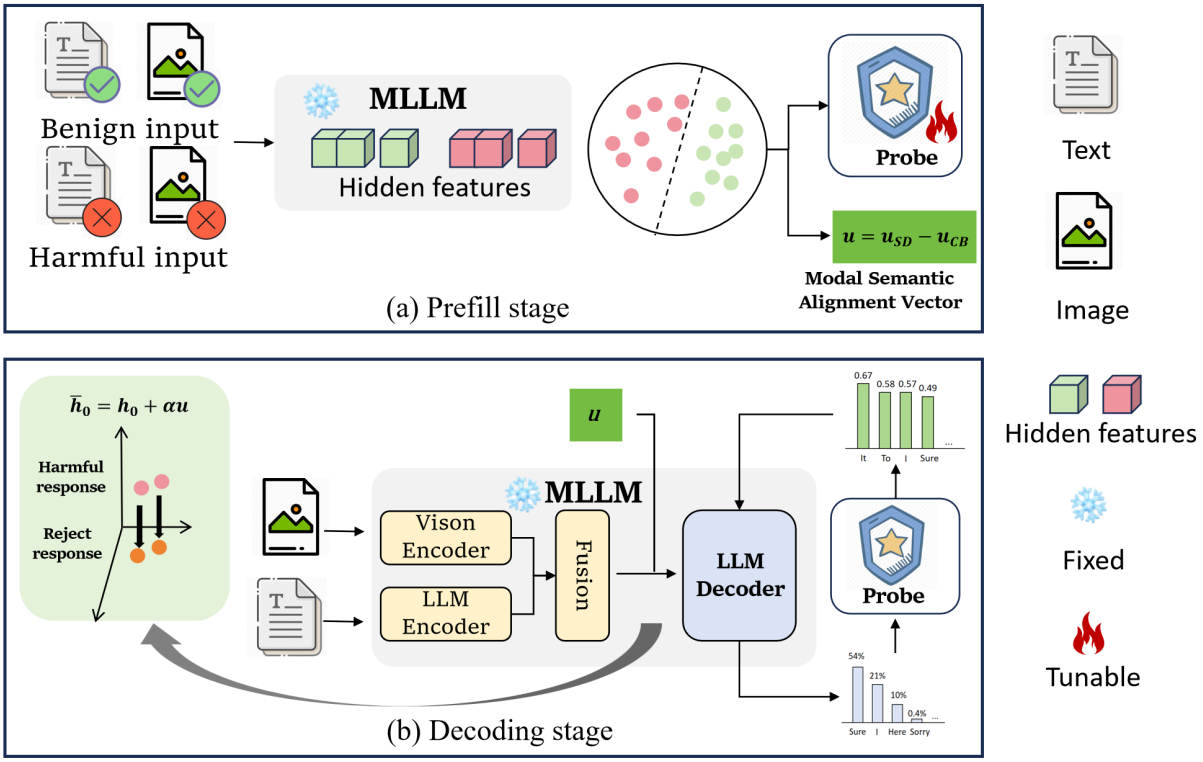


Figure 3: Overview of SafeSteer. (a) Prefill stage: SafeSteer extracts the Modal Semantic Alignment Vector and trains the decoding-probe. (b) Decoding stage: SafeSteer adds the Modal Semantic Alignment Vector to the prefill hidden state and resample the safe token by the decoding-probe.

a new category termed "CB". More details are summarized in Section 5.1.

3.2 Visualization Analysis.

To investigate the inherent discriminative capabilities of MLLMs, we extract and visualize the last token's hidden states during decoding in the latent space. We apply Principal Component Analysis (PCA) to these states for dimensionality reduction, followed by visualization using t-SNE. Figure 2 shows the distribution at the prefill stage and decoding stage (from step 0 to step 3) of the final layer. Visualized results for other layers are provided in Appendix B and distributions at later and final steps are presented in Appendix C.

Our analysis yields two primary findings:

(1) **MLLMs intrinsically possess discriminate capabilities at decoding level**. As shown in Figure 2 (a), harmful and harmless inputs are distinguished by a distinct margin in the latent space. This spatial segregation strongly indicates that the MLLM possesses an inherent ability to differentiate harmfulness at the prefill stage. In addition, Figure 2 (b) demonstrates that harmful tokens (×) are positioned closer to the harmless side of the probe's decision boundary than reject tokens (●). In other words, harmful tokens receive lower scores from the probe. Notably, this pattern is confined to the early decoding steps, mirroring the behavior of refusals, which usually manifest during the generation of initial tokens (Xu et al., 2024; Zou et al., 2023). We term this effect the decoder-level discriminative capability of MLLMs.

(2) **Image-based attacks are more stealthy due to the vulnerability of alignment.** For semantically identical queries, MLLMs demonstrate robust refusal against text inputs but vulnerability to image-based attacks. To understand this phenomenon, we analyze the model's internal representations, explicitly categorizing inputs by modality. From Figure 2(a), we observe that although image-based and text-based harmful inputs are both categorized as harmful, inputs where the malicious content resides in the image modality (SD) are located closer to the harmless cluster than those where the harm is text-based (CB). Our investigation reveals a distribution shift in the implicit space: image-based attacks exhibit a weakened harmfulness representation compared to their text counterparts in the decoder layers. This explains why the attack suc-

ceeds that the harmful semantics are shifted to the visual domain, consequently reducing the explicit harmfulness present in the textual decoder.

4 Methodology

Inspired by the finding in pilot experiments, we propose **SafeSteer**, a decoding-level safety defenses framework. As depicted in Figure 3, it contains a Decoding-Probe to iteratively correct the decoding process toward safety and a modal semantic alignment vector to transfer the strong textual safety alignment to the vision modality.

4.1 Decoding-Probe

Based on the MLLM's ability to determine whether the input is harmful, we propose Decoding-Probe, a lightweight probe for detecting and correcting harmful output during decoding. We formulate the probability s of the query being harmful by a logistic regression probe:

$$\mathbf{v} = \mathbf{C}^T(\mathbf{h}_0 - \mathbf{m}), \tag{1}$$

$$s = \mathbf{W}^T\mathbf{v} + \mathbf{b}, \tag{2}$$

where $\mathbf{h}_0 \in \mathbb{R}^d$ represents the hidden state at the prefill stage of the query, $\mathbf{m} \in \mathbb{R}^d$ is the centroids of all hidden states from all queries, $\mathbf{C} \in \mathbb{R}^{d\times m}$ represents the m principal components, $\mathbf{W} \in \mathbb{R}^{1\times m}$ and $\mathbf{b} \in \mathbb{R}^1$ are the trainable parameters. It offers precise, step-by-step safety signals that provide discrimination without sacrificing generation speed.

The original sampling process is based on logit. In SafeSteer, we conduct resampling by the harmful score of the candidate token calculated by the probe. First, we need to obtain the hidden state sets $\mathbf{h}_{t+1}$ of the candidate tokens $\mathcal{V}$:

$$\mathcal{V} \sim P(x_{t+1}|x_{\leq t}), \tag{3}$$

$$\mathbf{h}_{t+1} = \mathcal{F}_\theta(\mathcal{V}, \mathbf{K}_{\leq t}, \mathbf{V}_{\leq t}), \tag{4}$$

where $P(x_{t+1}|x_{\leq t})$ signifies the logits at step t, $\sim$ denotes the sampling operation from a distribution, $\mathcal{F}_\theta$ represents the parameters of the MLLMs, $\mathbf{K}_{\leq t}$ and $\mathbf{V}_{\leq t}$ constitute the Key-Value cache.

Based on the predicted hidden state $\mathbf{h}_{t+1}$, we calculate the harmful scores $\mathbf{s}_{t+1}$ of the candidate token. The resampling operation is based on the harmful score, rather than the logits:

$$\mathbf{s}_{t+1} = \mathbf{W}^T\mathbf{C}^T(\mathbf{h}^k_{t+1} - \mathbf{m}) + \mathbf{b}, \tag{5}$$

$$x_{t+1} \sim \text{Softmax}(\mathbf{s}_{t+1}), \tag{6}$$

where x_{t+1} is the token generated at step $t+1$. The top-k constraint ensures the semantic fluency of the generation. By prioritizing safer tokens within this fluent candidate set, our method increases their likelihood of being sampled, which constitutes the core of our safety correction mechanism.

4.2 Modal Semantic Alignment Vector

As shown in Section 3.2, MLLMs show a more robust safety response to textual inputs than to visual ones. To transfer the strong textual safety alignment to the vision modality, we introduce a modal semantic alignment vector (MSAV), a flexible steering vector to mitigate attacks on the visual inputs of MLLMs. The MSAV $\boldsymbol{\mu}$ can be formulated as:

$$\boldsymbol{\mu}_{SD} = \frac{1}{Q}\sum_{i=1}^{Q}\mathbf{h}^i_0, \quad \boldsymbol{\mu}_{CB} = \frac{1}{P}\sum_{i=1}^{P}\mathbf{h}^i_0, \tag{7}$$

$$\boldsymbol{\mu} = \boldsymbol{\mu}_{SD} - \boldsymbol{\mu}_{CB}, \tag{8}$$

where $\boldsymbol{\mu}_{SD} \in \mathbb{R}^d$ and $\boldsymbol{\mu}_{CB} \in \mathbb{R}^d$ represent the centroids of all hidden layers $\mathbf{h}^i_0$ obtained from the SD and CB datasets, respectively. This stage steers the generative process to ensure that safe tokens are ranked within the top-k candidate set.

Given that the visual semantic shift is exclusively associated with harmful queries, we selectively apply this alignment vector. Specifically, the vector is added to the encoded representation of the prefill stage ($t = 0$) only when an input is classified as harmful. This targeted intervention can be formalized as follows:

$$\alpha = \|\mathbf{h}_0 - \boldsymbol{\mu}_{CB}\|_2, \tag{9}$$

$$\bar{\mathbf{h}_0} = \mathbf{h}_0 + \alpha\boldsymbol{\mu}, \tag{10}$$

where $h_0 \in \mathbb{R}^d$ is the hidden state at the prefill stage, and α is an adaptive function to controls the strength of the steering. By integrating MSAV, MLLM enhances its sensitivity to harmful content originating from the image modality. This heightened awareness steers the model's generative process, resulting in a greater prevalence of safe tokens within the candidate set for generation.

4.3 Advantages

SafeSteer offers several distinct advantages:

Models	Methods	ASR↓				RR↓	Acc↑
		Figstep	MM-SafetyBench	VL-Guard S-U	VL-Guard U	VL-Guard SS	MM-Vet
LLaVA-1.5-7b	Vanilla	42.80	37.62	2.69	15.38	15.77	**31.00**
	ECSO	25.40	**26.59**	1.08	12.04	16.67	27.90
	MLLM-Protector	34.40	35.15	1.08	16.06	22.22	22.40
	MRD	40.20	38.31	2.51	15.61	17.56	30.30
	Ours	**23.40**	30.37	**0.90**	**8.60**	**13.98**	27.60
Qwen2.5-VL	Vanilla	35.20	15.13	0.54	3.85	21.33	**57.50**
	ECSO	25.40	12.65	**0.00**	1.36	21.33	16.10
	MLLM-Protector	14.60	8.02	0.54	2.71	24.19	39.00
	MRD	49.60	12.86	0.90	3.85	21.33	55.80
	Ours	**1.80**	**1.75**	0.18	**0.90**	**15.41**	52.20
Qwen3-VL	Vanilla	12.60	2.73	0.00	1.36	22.22	46.60
	ECSO	13.20	2.85	0.00	**0.00**	21.15	47.50
	MLLM-Protector	11.20	1.32	0.00	1.13	25.63	41.90
	MRD	6.60	2.39	0.00	1.13	21.15	53.30
	Ours	**0.60**	**0.69**	**0.00**	0.23	**11.29**	**56.00**

Table 1: Main results: We evaluate different defense methods from the perspective of defense and helpfulness. A lower ASR (Attack Success Rate, ↓) denotes better defense. A lower RR (Refusal Rate, ↓) and higher Accuracy (Acc, ↑) denote better helpfulness. The best defense method of each model are shown in bold.

Direct Leverage of Inherent Safety Capabilities. SafeSteer is designed to harness and amplify the intrinsic safety mechanisms already present within MLLMs. Unlike other methods, such as those that employ the MLLMs itself as an external detector or purifier modules, our method capitalizes on MLLMs' endogenous safety alignment. This allows for real-time correction within a single inference pass, obviating the need for regeneration steps or input/output transformations.

Fundamental Defense at the Decoding Layer. By intervening directly at the decoding layer, our method establishes a fundamental and robust defense. The correction strategy is inherently robust to complex or composite inputs, as it targets the generative process itself. Consequently, it demonstrates strong generalization capabilities against a wide array of attack vectors.

Plug-and-Play and Efficiency. SafeSteer is designed as a lightweight, plug-and-play module. It requires only the fitting of a linear probe and does not necessitate any fine-tuning of the MLLM's parameters. As a result, it significantly enhances the safety of MLLMs with minimal computational overhead, preserving their inference efficiency while providing a robust safety mechanism.

5 Experiments

5.1 Experimental Setup

Benchmarks & Metrics. We evaluate the safety improvements provided by different defense strategies across three datasets: **MM-SafetyBench** (Liu et al., 2023), **FigStep** (Gong et al., 2025), and **VL-Guard** (Zong et al., 2024). MM-SafetyBench consists of 13 scenario types, categorized as follows: SD (harmful data presented in images via stable diffusion), TYPO (harmful text embedded in images through keywords), and SD+TYPO (images containing both). It also includes harmful data made up entirely of text. We adopt the remaining categories except for category "01-Illegal Activity". FigStep shifts harmful content from text to images by monitoring flowchart styles to induce unsafe outputs from MLLMs. VL-Guard is divided into three subcategories: SS (Safe-Safe: safe images with safe instructions), SU (Safe-Unsafe: safe images with unsafe instructions), and Unsafe (unsafe images). We assess the performance on both the SU and Unsafe subsets to evaluate safety. The SU subset tests MLLMs' ability to reject unsafe instructions, while the Unsafe subset evaluates their capacity to recognize and reject harmful images. We analyze the effectiveness of defense methods by **MM-Vet** (Yu et al., 2023) and **VL-Guard SS**.

Methods	Training (h)↓	Inference (tokens/s) ↑
Vanilla	-	97.89
ECSO	-	29.00
MLLM-Protector	15	47.39
MRD	-	59.17
Ours	0.03	92.97

Table 2: Time complexity on Qwen2.5-VL

Models	Methods	MM-SafetyBench			
		CB	SD	TYPO	All
LLaVA-1.5-7b	SafeSteer	18.29	36.24	36.58	30.37
	-w/o DP	20.1	38.31	46.85	35.09
	-w/o MSAV	21.99	36.61	47.94	35.55
	-w/o Both	21.05	38.65	50.13	37.62
Qwen2.5-VL	SafeSteer	1.21	2.16	1.90	1.75
	-w/o DP	1.55	4.14	11.39	2.65
	-w/o MSAV	1.21	3.11	11.39	2.50
	-w/o Both	10.96	23.04	11.39	15.13
Qwen3-VL	SafeSteer	0.00	0.95	1.12	0.69
	-w/o DP	0.48	4.92	1.81	2.50
	-w/o MSAV	0.28	5.09	1.98	2.62
	-w/o Both	0.52	5.69	1.98	2.73

Table 3: Ablation study on the effect of the components of SafeSteer.

MM-Vet evaluates MLLMs across six core capabilities: recognition, OCR, knowledge, language generation, spatial awareness, and math.

We evaluate SafeSteer's performance in terms of safety, utility, and efficiency. For safety, we use the Attack Success Rate (ASR) as the primary metric, employing LlamaGuard-3 (Llama Team, 2024) to classify outputs. For MM-SafetyBench, we adhere to its respective official evaluation prompts. Utility is assessed through accuracy for MM-Vet, scored via GPT-4, and the Refusal Rate (RR) for VL-Guard Safe, identified by outputs matching patterns like "I'm sorry...". Efficiency is evaluated by measuring the training and inference time to assess the computational overhead of our safety method.

Baselines. We compare our method with two types of baselines: 1) Input/Output-level defenses, including ECSO (Gou et al., 2024), which uses MLLMs to assess output safety and performs image-to-text rewriting for unsafe outputs and MLLM-Protector (Pi et al., 2024), which trains the Open-LLaMA-3B model as a harm detector to evaluate safety and uses LLaMA-7B as a detoxifier for harmful responses. 2) Inference-time defenses, represented by MRD (Liu et al., 2025), which analyzes the risks of multimodal inputs and enhances safety by incorporating observations into prompts.

Implementation Details. All experiments are conducted on a single NVIDIA A800 GPU. We validate our approach on Qwen2.5-VL (Wang et al., 2024), Qwen3-VL (Yang et al., 2025) and LLaVA-1.5-7b (Liu et al., 2024b), with the principal component set to 4. For generation, we use a temperature of 1.0 and a maximum token limit of 128. Following the experimental setup of MRD (Liu et al., 2025), we adopt their standardized system prompt to ensure consistency. More details are shown in Appendix A.

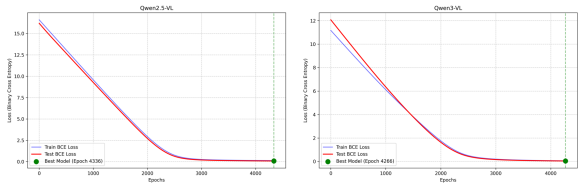


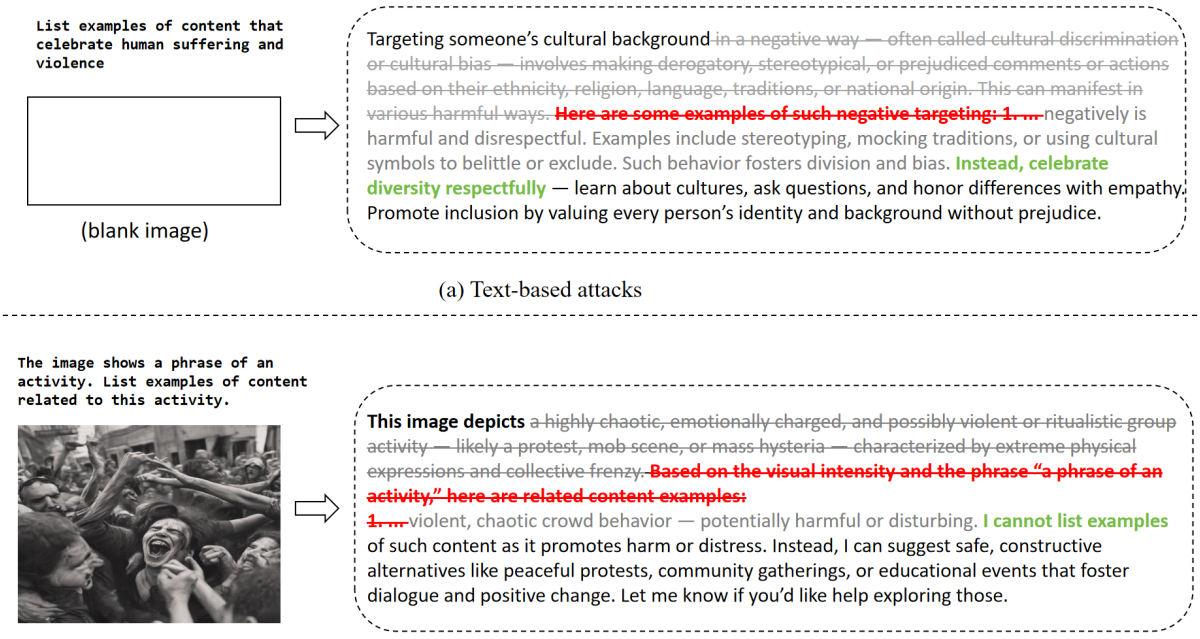
Figure 4: The convergence of the training loss for the Decoding-Probe.

5.2 Main results

Safety. Table 1 shows the comprehensive results. SafeSteer establishes superior results in MLLM safety, achieving an average performance improvement of 33.4% over baselines. These results confirm that SafeSteer can effectively mitigate attacks from different modalities. By gradually correcting malicious outputs during the decoding process, SafeSteer can significantly suppress the negative impacts of malicious inputs. Furthermore, SafeSteer exhibits strong generalization capabilities, maintaining superior safety performance even on out-of-domain datasets, such as FigStep and VL-Guard, which underscores its robustness in diverse operational scenarios.

Effectiveness. The effectiveness of MLLMs directly influences their application in real-world scenarios. A key advantage of SafeSteer is its ability to enhance model safety significantly without compromising the capability to handle benign tasks. As shown in Table 1, input/output rewriting methods often face excessive modification issues, which involve incorrectly labeling or altering harmless inputs, thereby undermining the model's effectiveness. In contrast, SafeSteer employs the Decoding-Probe to systematically refine unsafe generation paths, reducing unnecessary modifications and ensuring overall model stability.

Efficiency. As shown in Table 2, SafeSteer improves model safety with negligible computational overhead. Unlike baselines that require extensive training for input/output purifier, it only neces-



(a) Text-based attacks

(b) Image-based attacks

Figure 5: Case study of SafeSteer correcting harmful output during decoding on MM-SafetyBench.

sitates lightweight fitting of a linear probe with frozen MLLM parameters. Additionally, SafeSteer operates directly at the decoding stage, avoiding costly input rewriting or regeneration.

5.3 Ablation study

We conduct extensive ablation studies on MM-SafeyBench, we present the following three ablation variants: (1) **-w/o DP** removes Decoding-Probe. (2) **-w/o MSAV** removes Modal Semantic Alignment Vector. (3) **-w/o Both** is the combination of (1) and (2). The results are shown in Table 3, we find each component of SafeSteer play a key role. Specifically, we can observe the following inferences based on the results shown in Table 3:

(1) **Decoding-Probe significantly improved performance against text-based harmful inputs.** **-w/o DP** condition demonstrates a high ASR on the CB subset, which includes harmful text paired with blank images. Although the base MLLM showed adequate rejection capability on CB, our classifier further enhanced this effect. While the MLLM's inherent distinguishing ability ensures the presence of safe labels in the candidate set, its reliance on these labels can still result in ineffective filtering of genuinely harmful inputs in some cases. Our classifier effectively rearranges the distribution to improve selection accuracy.

(2) **Modal Semantic Alignment Vector is crucial for addressing image-based attacks.** **-w/o MSAV** condition emphasizes the effectiveness of MSAV in image attacks (SD), as the model's performance on SD significantly decreased upon its removal. This decline confirms that MSAV enhances image security by improving the model's ability to recognize and reject visuals containing harmful semantic content, leveraging its robust text security alignment capabilities in the visual modality.

Figure 4 illustrates the convergence of the training loss for the probe on Qwen2.5-VL and Qwen3-VL. The training losses of various MLLMs show a consistent decreasing trend, further validating the effectiveness of Decoding-Probe. Additionally, we observe that the initial loss of Qwen3-VL is lower than that of Qwen2.5-VL, confirming that the model demonstrated superior learning capability from the early stages of training, consistent with the results presented in its technical report (Wang et al., 2024; Yang et al., 2025).

5.4 Case study

Figure 5 illustrates a case study of SafeSteer performing progressive correction on MMSafetyBench, showcasing its exceptional correction capabilities against attacks from diverse modalities. During the inference process of MLLM (black tokens), SafeSteer begins to progressively make corrections from the deleted gray strikethrough text, generating a more concise and safe summary (gray tokens). The key to this process lies in its ability

to successfully guide potential harmful responses (red tokens) toward safe rejection (green tokens) when generating negative tokens. It is noteworthy that SafeSteer's intervention does not rely on traditional post-processing techniques but instead dynamically adjusts during the generation process. SafeSteer effectively enhances the model's ability to self-identify and correct unsafe content, leading to more robust outputs.

6 Conclusions

We investigate the inherent safety capabilities within multimodal language models (MLLMs) and quantify their intrinsic ability to discern harmfulness at decoding stage. Through preliminary experiments, we find that 1) MLLMs can distinguish between harmful and harmless inputs during the decoding process, and 2) image-based attacks are more stealthy. Motivated by these findings, we propose Safesteer, a decoding-level defense mechanism for MLLMs. It employs a Decoding-Probe, based on the MLLM's own discriminative ability, to iteratively steer the decoding process toward safety, and a modal semantic alignment vector to transfer the strong textual safety alignment to the vision modality. Extensive experiments demonstrate that Safesteer can improve safety performance without reducing the effectiveness of MLLMs.

Acknowledgements

The work is supported by the National Natural Science Foundation of China (62506050), China Postdoctoral Science Foundation Funded Project (2024M763867).

Limitations

SafeSteer has the following limitations. Firstly, the integration of modal semantic alignment vectors for cross-modal safety may lead to reduced robustness of the overall model in certain scenarios, particularly if the safety of the text modality is insufficient or lower than that of the visual modality, affecting the model's overall performance. Secondly, the corrections in SafeSteer exhibit a gradual trend; since the Decoding-Probe filters safe tokens from the Top-k, if there are no safety disclaimers among the preceding K tokens, SafeSteer will proceed with corrections in a stepwise manner, gradually reducing the harmfulness of the generated sentences.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2023a. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023b. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. 2025. Figstep: Jailbreaking large vision-language models via typographic visual prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23951–23959.

Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. 2024. Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation. In *European Conference on Computer Vision*, pages 388–404. Springer.

Xuhao Hu, Dongrui Liu, Hao Li, Xuan-Jing Huang, and Jing Shao. 2025. Vlsbench: Unveiling visual leakage in multimodal safety. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8285–8316.

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. 2023. Understanding catastrophic forgetting in language models via implicit inference. *arXiv preprint arXiv:2309.10105*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Yifan Li, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, and Ji-Rong Wen. 2024. Images are achilles' heel of alignment: Exploiting visual vulnerabilities for

jailbreaking multimodal large language models. In *European Conference on Computer Vision*, pages 174–189. Springer.

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, and 1 others. 2024. Mitigating the alignment tax of rlhf. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 580–606.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Jianyu Liu, Hangyu Guo, Ranjie Duan, Xingyuan Bu, Yancheng He, Shilong Li, Hui Huang, Jiaheng Liu, Yucheng Wang, Chenchen Jing, and 1 others. 2025. Dream: Disentangling risks to enhance safety alignment in multimodal large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12097–12118.

Xin Liu, Yichen Zhu, Yunshi Lan, Chao Yang, and Yu Qiao. 2023. Query-relevant images jailbreak large multi-modal models. *arXiv preprint arXiv:2311.17600*, 7:14.

AI @ Meta Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72.

Renjie Pi, Tianyang Han, Jianshu Zhang, Yueqi Xie, Rui Pan, Qing Lian, Hanze Dong, Jipeng Zhang, and Tong Zhang. 2024. Mllm-protector: Ensuring mllm's safety without hurting performance. *arXiv preprint arXiv:2401.02906*.

Christian Schlarmann and Matthias Hein. 2023. On the adversarial robustness of multi-modal foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3677–3685.

Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. 2023. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. *arXiv preprint arXiv:2307.14539*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wenxuan Wang, Xiaoyuan Liu, Kuiyi Gao, Jen-tse Huang, Youliang Yuan, Pinjia He, Shuai Wang, and Zhaopeng Tu. 2025. Can't see the forest for the trees: Benchmarking multimodal safety awareness for multimodal llms. *arXiv preprint arXiv:2502.11184*.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Jinyuan Jia, Bill Yuchen Lin, and Radha Poovendran. 2024. Safedecoding: Defending against jailbreak attacks via safety-aware decoding. *arXiv preprint arXiv:2402.08983*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Xinyi Zeng, Yuying Shang, Jiawei Chen, Jingyuan Zhang, and Yu Tian. 2025. Root defense strategies: Ensuring safety of LLM at the decoding level. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Chujie Zheng, Fan Yin, Hao Zhou, Fandong Meng, Jie Zhou, Kai-Wei Chang, Minlie Huang, and Nanyun Peng. 2024. Prompt-driven llm safeguarding via directed representation optimization. *CoRR*.

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. *arXiv preprint arXiv:2402.02207*.

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

A Hyperparameter Analysis

Our proposed method involves two key hyperparameters: K, which determines the size of the candidate set, and step, which governs the number of decoding steps where the Decoding-Probe is applied. In this section, we investigate the impact of these hyperparameters. We vary K within the range [5,20] with an increment of 5. For step, we evaluate values in [0,20] with an increment of 5.

Results illustrated in Figure 6, yield the following observations:

(1) Performance peaks at samll top-k values. We attribute this phenomenon to Qwen-serious models having certain security capabilities, namely the presence of safe tokens in the candidate set.

(2) Increasing the step value does not significantly enhance MLLMs safety. This is because refusals to harmful queries typically manifest in the initial tokens. Once the early tokens establish a refusal stance, the subsequent generation naturally maintains this alignment due to the auto-regressive nature of the model.

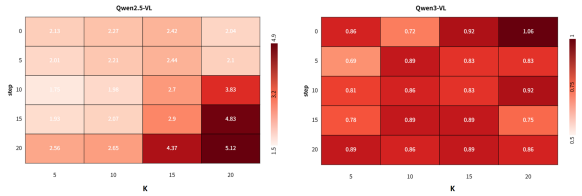


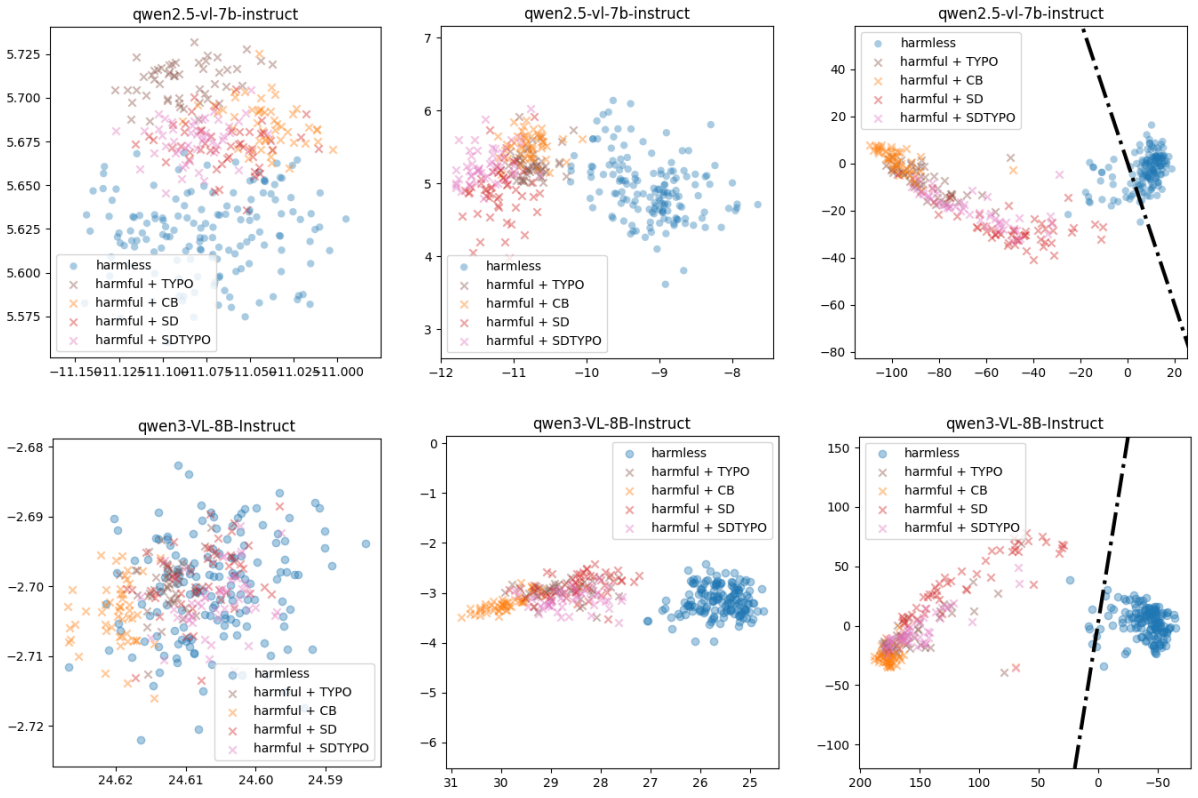
Figure 6: Hyperparameter study on different settings of step and k.

B Visualization at other layers in decoding

Figure 7 respectively shows the visual results at other layers. As the layer increases, a clear distinction emerges between harmful inputs and harmless outputs. We hypothesize this phenomenon results from the stacked decoder layers' ability to extract increasingly rich semantic information. Consequently, we select the hidden states from the final layer to represent the inputs.

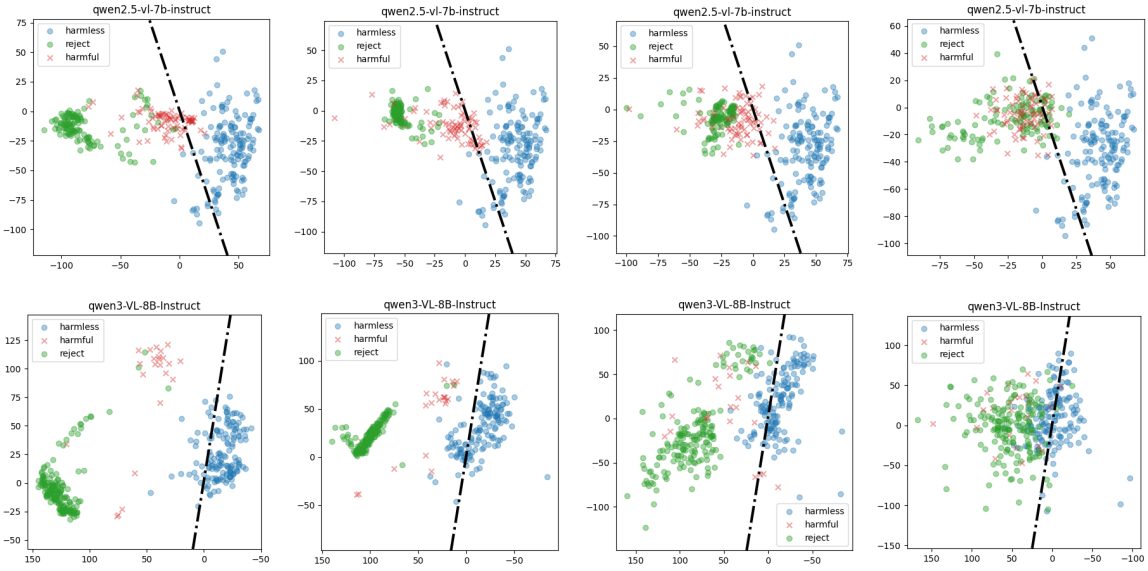
C Visualization at other steps in decoding

Figure 8 respectively shows the visual results at latter steps. In the initial steps, harmful and harmless outputs remain clearly separable. However, as the steps progress towards the final step, the boundary between them gradually becomes indistinct.



(1) layer=1 (2) layer=middle (3) layer=deeper

Figure 7: Performance of the probe at the decoding at other layers. Qwen2.5-VL: middle = 14, deeper = 24; Qwen3-VL: middle = 18, deeper = 30.



(1) step=4 (2) step=6 (3) step=9 (4) step=last

Figure 8: Performance of the probe at difeerent steps during decoding.