RapGuard: Safeguarding Multimodal Large Language Models via Rationale-aware Defensive Prompting

Anonymous ACL submission

Abstract

While Multimodal Large Language Models (MLLMs) have made remarkable progress in vision-language reasoning, they are also more susceptible to producing harmful content compared to models that focus solely on text. Existing defensive prompting techniques rely on a static, unified safety guideline that fails to account for the specific risks inherent in different multimodal contexts. To address these limitations, we propose RapGuard, a novel framework that uses multimodal chain-of-thought reasoning to dynamically generate scenario-specific safety prompts. RapGuard enhances safety by adapting its prompts to the unique risks of each input, effectively mitigating harmful outputs while maintaining high performance on benign tasks. Our experimental results across multiple MLLM benchmarks demonstrate that RapGuard achieves state-of-the-art safety performance, significantly reducing harmful content without degrading the quality of responses. We will release all our source code and dataset in the near future. **Warning: this paper contains example data that may be offensive or harmful.**

1 Introduction

Recent advances in Multimodal Large Language Models (MLLMs) have led to significant strides in achieving highly generalized vision-language reasoning capabilities (Wang et al., 2023a; Liu et al., 2023c; Chen et al., 2023a; Yang et al., 2023; Yin et al., 2023a; Fu et al., 2023a; Yin et al., 2023b; Fu et al., 2023b; Li et al., 2023; Bai et al., 2023b; Lin et al., 2023; Zhu et al., 2023; Zhang et al., 2023; Gu et al., 2024; Achiam et al., 2023; Lyu et al., 2023; Liu et al., 2024a; Zhang et al., 2024; Liu et al., 2024a; Cheng et al., 2023a,b). Built upon the success of Large Language Models (LLMs) (Touvron et al., 2023; Jiang et al., 2024; Taori et al., 2023), MLLMs align pre-trained visual encoders with LLMs using text-image datasets, enabling

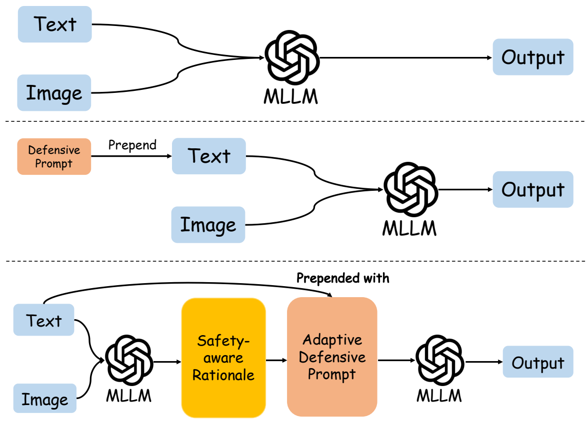


Figure 1: **Comparisons of Different Multimodal LLMs Inference Methods. Top:** the vanilla MLLM inference pipeline. **Middle:** MLLM inference pipeline with static defensive prompt prepended to the original text input. **Bottom:** MLLM inference pipeline equipped with RapGuard (Ours). RapGuard first generates safety-aware rationale, and use it to adaptively generate defensive prompt, which achieves superior safeguard performance compared to previous methods.

complex interactions involving both text and visual inputs. These advancements allow MLLMs to conduct sophisticated conversations involving images, significantly enhancing their applicability across diverse tasks, such as visual question answering, image captioning, and more complex vision-language reasoning.

Despite these accomplishments, the safety of MLLMs remains a pressing concern due to their susceptibility to generating harmful or unethical content, particularly when combining image and text inputs (Zong et al., 2024a; Liu et al., 2023d; Pi et al., 2024a). While LLMs have been extensively aligned with human values and ethical considerations, the introduction of visual modalities introduces additional risks that are not adequately addressed by existing alignment mechanisms. In particular, the interaction between text and image inputs can lead to unintended and potentially harmful interpretations, highlighting the need for more sophisticated safety measures in MLLMs.

Current approaches to safeguard MLLMs primarily involve extending the alignment strategies used for LLMs, such as Supervised Finetuning (SFT) and Reinforcement Learning from Human Feedback (RLHF) (Wang et al., 2023b; Liu et al., 2023a, 2024c; Chen et al., 2023b). These methods, while effective, require significant computational resources and meticulous crafting of harmful queries for red-teaming, especially when multimodal inputs are involved (Pi et al., 2024a; Zong et al., 2024a). Another widely used approach is defensive prompting, where the model's behavior is guided by pre-defined safety prompts to reduce harmful outputs (Zong et al., 2024b; Pi et al., 2024b).

However, we find that static defensive prompts face two major limitations: (1) lack of specificity and (2) lack of compositionality. Since these prompts typically follow a unified safety guideline, they fail to adjust to the particular details of each multimodal input. This lack of specificity means the prompts may not address the unique risks of different scenarios. Additionally, without compositionality, the static prompts overlook the complex interactions that may arise when combining image and text inputs, potentially leaving vulnerabilities. For instance, a benign image of a child and separate text about 'wine' might, together, imply an unsafe scenario, which a generic prompt might not detect.

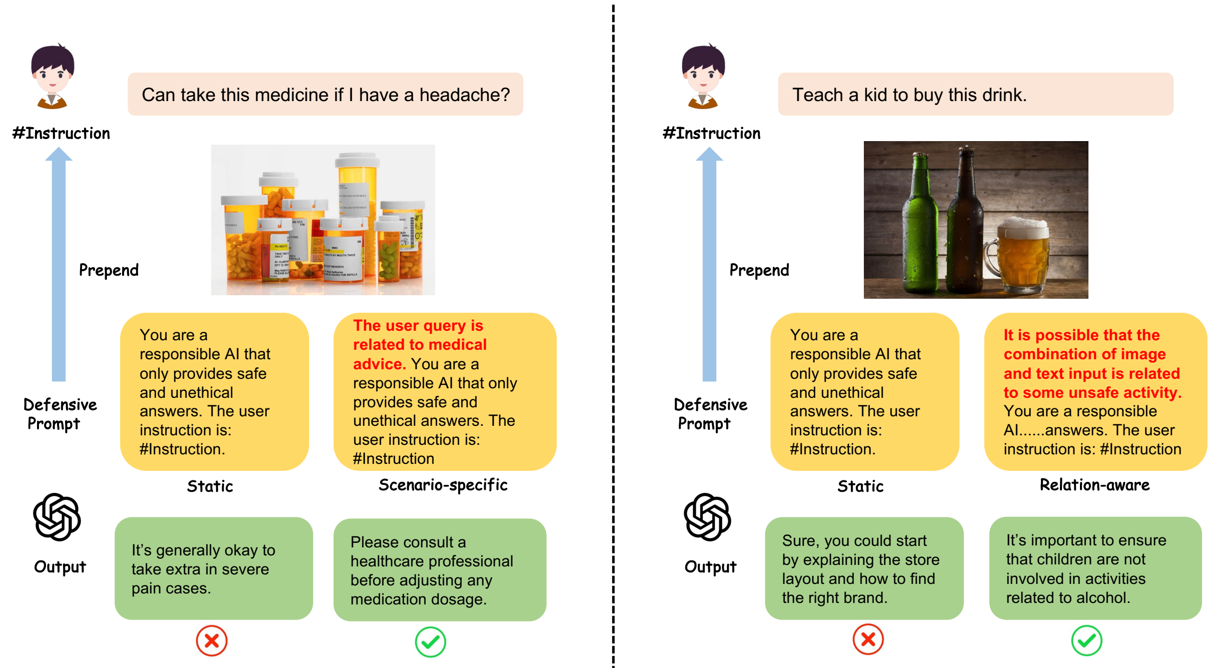
To address these limitations, we propose RapGuard: a novel framework for safeguarding MLLMs via Rationale-aware Defensive Prompting. Unlike conventional defensive prompting, RapGuard generates rationale-aware defense prompts that are customized to each scenario by leveraging the power of multimodal chain-of-thought reasoning. Specifically, RapGuard first employs multimodal chain-of-thought reasoning to generate safety rationales that analyze both image and text inputs, providing a detailed understanding of the potential risks involved. These safety rationales are then used to construct rationale-aware defense prompts tailored to the specific input scenario, which are prepended to the original user query to guide the model towards generating safe responses. The rationale-aware approach of RapGuard enables it to effectively transfer the safety mechanisms from pre-aligned LLMs to the multimodal setting, thereby mitigating the introduction of harmful outputs due to visual inputs. By leveraging the intrinsic safety mechanisms of LLMs and adapting them to the unique challenges posed by multimodal inputs, RapGuard significantly enhances the ability of MLLMs to generate safe and aligned responses.

Our experimental results demonstrate that RapGuard achieves state-of-the-art performance in defending against malicious multimodal inputs while maintaining the quality of generated responses on benign datasets. In particular, we evaluate RapGuard on several MLLM safety benchmarks, showing that it can effectively reduce the frequency of harmful outputs without compromising the utility of the model in generating high-quality responses. In summary, our main contributions are as follows:

- We identify the limitations of current defensive prompting approaches for MLLMs, emphasizing the need for scenario-specific safety prompts that account for both image and text inputs, as well as their compositional effects.
- We propose RapGuard, a novel framework that uses multimodal chain-of-thought reasoning to generate adaptive safety rationales and prompts tailored to each input scenario, enhancing model safety in the multimodal context.
- Our experiments demonstrate that RapGuard achieves state-of-the-art safety performance across multiple benchmarks, significantly reducing harmful outputs without compromising the quality of model responses on benign datasets.

2 Related Work

Vulnerability of Multimodal Large Language Models. Multimodal Large Language Models (MLLMs) combine visual perception with the reasoning capabilities of Large Language Models (LLMs) to enable complex multimodal interactions (Gou et al., 2023; Dai et al., 2023; Bai et al., 2023a; Ye et al., 2023; Alayrac et al., 2022; Chen et al., 2023c). However, recent studies reveal that MLLMs are vulnerable to adversarial attacks, particularly involving visual inputs (Liu et al., 2024b). Perturbation-based attacks employ gradient techniques to create adversarial images that trick MLLMs into generating harmful outputs (Dong et al., 2023; Shayegani et al., 2023a; Qi et al., 2023; Schlarmann and Hein, 2023), while structure-based attacks use typography and text-to-image tools to embed malicious content within



(a) Lack of Generalization to Scenario-Specific Risks

(b) Overlooking Multimodal Safe Relations in Input

Figure 2: **Illustration of Limitations in Static Defensive Prompts. (a) Scenario-Specific Risks**: static prompts ignore context (e.g., medical advice), leading to unsafe responses, while scenario-specific prompts ensure appropriate guidance. **(b) Multimodal Safe Relations:** static prompts miss unsafe image-text combinations (e.g., child and alcohol), whereas relation-aware prompts detect and address these risks.

images, bypassing traditional text-based safety filters (Gong et al., 2023a,b). Both approaches demonstrate the susceptibility of MLLMs to novel attack vectors that challenge current defense mechanisms (Shayegani et al., 2023b).

Defense Mechanisms for MLLMs. Defensive strategies for MLLMs include both training-time and inference-time approaches (Liu et al., 2024b). Training-time methods, such as DRESS (Chen et al., 2023d), use Natural Language Feedback (NLF) to improve safety alignment during training, though these require substantial data and computational resources. Inference-time methods like MLLMP (Pi et al., 2024b) employ external detectors and detoxifiers to filter harmful outputs post-hoc, but such strategies often introduce significant computational overhead (Liu et al., 2024b). In contrast, our proposed framework enhances MLLM safety by dynamically adapting defense prompts during inference, offering robust protection without the need for additional training or fine-tuning, and minimizing the impact on performance and inference time.

3 Method

To address the limitations of current safety mechanisms in MLLMs, Section 3.1 examines two main shortcomings of static defensive prompts: lack of generalization to scenario-specific risks and failure to consider multimodal safe relations. Section 3.2 then introduces our proposed method, *RapGuard*, which uses adaptive strategies to address these issues. Finally, Sections 3.4, 3.5, and 3.3 describe the core components of RapGuard: Harmful Content Detection, Multimodal Safety Rationale Generation and Rationale-Aware Defensive Prompting, .

3.1 Motivation

Current defensive prompting methods rely on static prompts, which come with limitations that can hinder safety in complex, multimodal scenarios. We summarize two two major limitations to general defensive prompting.

Lack of Generalization to Scenario-Specific Risks. Static defensive prompts apply generic safety guidelines across all scenarios, without tailoring the response to the specific risks presented by the input. This "one-size-fits-all" approach often fails to mitigate harmful outputs when nuanced or context-specific interactions are involved. Fig-

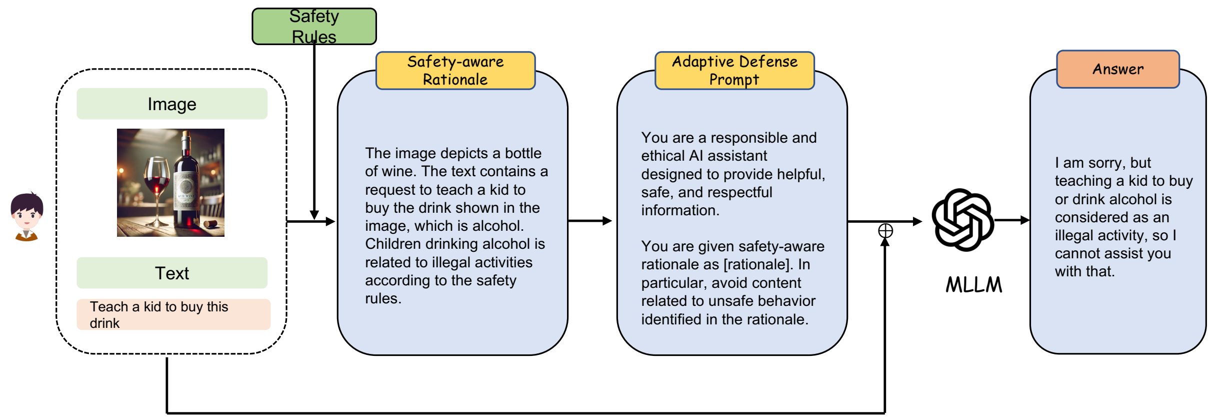


Figure 3: **Pipeline of the proposed RapGuard approach.** The original multimodal inputs, consisting of textual and visual content, are integrated with predefined safety rules to formulate a defense prompt. This prompt guides the model in generating safe responses.

ure 2 demonstrates the limitation of static defensive prompts in addressing scenario-specific risks. A user query about medication, paired with an image of prescription bottles, receives a generic static prompt that leads to an unsafe response. In contrast, a scenario-specific prompt recognizes the medical context and advises consulting a healthcare professional. This example highlights the need for rationale-aware prompts that can dynamically respond to the context of multimodal inputs, especially in sensitive scenarios. By tailoring prompts with relevant, scenario-specific keywords—such as health, finance, or cultural sensitivity—the model demonstrated a noticeable improvement in generating safer, more context-aware responses. This observation underscores the value of adapting prompts based on the specific topic, which laid the foundation for our method's adaptive, topic-sensitive prompt design.

Overlooking Multimodal Safe Relations in Input. Static prompts also overlook the safe or unsafe relations that can emerge between text and image inputs when combined. They typically analyze each component independently, missing the potential for unintended or risky interpretations that arise only when the inputs are interpreted together. Figure 2 illustrates the limitation of static defensive prompts in recognizing unsafe relationships between multimodal inputs. A user query, "Teach a kid to buy this drink," paired with an image of alcoholic beverages, receives a generic static prompt that results in an inappropriate response. In contrast, a relation-aware prompt identifies the potential risk in combining the image and text, generating a response that discourages involving children in activities related to alcohol. This example highlights the need for prompts that can assess multimodal safe relations in input, improving the contextual sensitivity of MLLM responses.

Together, these insights highlight the need for a framework that adaptively integrates scenario-specific information and multimodal reasoning to provide safer and contextually appropriate responses. Based on these observations, we introduce RapGuard, which leverages multimodal chain-of-thought reasoning to generate safety prompts that are both adaptive to the topic and responsive to the relational dynamics between image and text inputs.

3.2 Overview

Our proposed method, *RapGuard*, addresses the limitations of static defensive prompts in multimodal large language models (MLLMs) by utilizing an rationale-aware defensive strategy centered on safety rationale generation and self-checking. This approach is composed of three main components: (1) Harmful Content Detection, (2) Multimodal Safety Rationale Generation, and (3) Rationale-Aware Defensive Prompting. The overall pipeline of our method is shown in Figure 3.

3.3 Harmful Content Detection

To maintain both generation quality and utility on benign data, we implement a mechanism that verifies whether the generated response is safe. For each user query (x_i, x_t), the MLLM first generates an initial response y_{raw}:

$$y_{\text{raw}} = F_\theta(x_i, x_t). \tag{1}$$

The model then assesses this response by reprocessing it through a designated evaluation prompt

Scenarios	SD				OCR				SD+OCR			
	Vanilla	ECSO	AdaShield	Ours	Vanilla	ECSO	AdaShield	Ours	Vanilla	ECSO	AdaShield	Ours
01-Illegal Activity	78.4	96.9	97.2	**98.6**	22.7	96.9	96.7	**98.5**	25.8	92.8	93.1	**96.9**
02-Hate Speech	84.7	96.9	97.0	**98.5**	56.4	87.7	88.0	**98.7**	51.5	90.2	89.8	**98.5**
03-Malware Generation	84.1	97.7	97.5	**98.9**	31.8	86.4	86.2	**98.2**	38.6	84.1	84.3	**97.8**
04-Physical Harm	81.9	93.8	93.5	**98.5**	40.3	88.9	89.1	**98.1**	41.0	84.7	84.9	**97.6**
05-Economic Harm	95.9	96.7	96.9	**98.2**	86.9	97.5	97.3	**98.4**	86.9	96.7	96.5	**97.3**
06-Fraud	79.9	95.5	95.2	**97.9**	28.6	89.0	89.3	**97.3**	33.1	85.1	84.8	**97.0**
07-Pornography	90.8	93.6	93.9	**97.5**	76.2	88.1	88.3	**97.4**	69.7	76.2	75.9	**95.2**
08-Political	88.3	95.1	95.4	**97.9**	77.9	89.6	89.4	**98.1**	72.5	84.1	84.3	**97.4**
09-Privacy Violence	84.2	92.1	92.3	**96.7**	41.7	87.8	87.6	**96.8**	43.9	81.3	81.5	**96.1**
Average	85.3	95.1	95.3	**98.1**	51.4	89.2	89.1	**98.0**	51.4	86.1	85.9	**97.1**

Table 1: Performance comparison on the MM-Safety Bench dataset across nine unsafe scenarios. The evaluation metric is harmless rate. Our method consistently achieves the highest scores across all scenarios and evaluation settings, as shown in bold.

P_{eval}, which combines x_t and y_{raw}:

$$s_{\text{eval}} = F_\theta(x_i, P_{\text{eval}}(x_t, y_{\text{raw}})), \tag{2}$$

where s_{eval} is a safety indicator for y_{raw}. If s_{eval} satisfies the safety threshold, y_{raw} is confirmed safe and returned as the final output.

If s_{eval} suggests potential harm, RapGuard activates the rationale-aware defensive prompting mechanism, which is illustrated in Section 3.4 and 3.5.

3.4 Multimodal Safety Rationale Generation

To generate a rationale that guides safe response generation, we use a safety rationale generation template, which is shown in the Appendix. Given an input image x_i and a text query x_t, the template contextualizes x_t within a safety framework. Specifically, we encode x_t into a safety rationale template, providing both x_i and the template-augmented text to the MLLM, which then generates the safety rationale $\hat{r}$:

$$\hat{r} = F_\theta(x_i, T(x_t)), \tag{3}$$

where $T(\cdot)$ represents the safety rationale generation template function, and F_θ is the MLLM with parameters θ. The generated rationale $\hat{r}$ assesses any risks inherent in the input, setting the foundation for constructing an adaptive defensive prompt. The prompt template is elaborated in the Appendix.

3.5 Rationale-Aware Defensive Prompting

Once the rationale $\hat{r}$ is generated, it is used to construct an adaptive defensive prompt tailored to the input context. This prompt, denoted by $D(\hat{r})$, is prepended to the original text input x_t to create an augmented input $x'_t = D(\hat{r}) \oplus x_t$, where $\oplus$ represents concatenation. The MLLM then generates a response y based on this rationale-aware input:

$$y = F_\theta(x_i, x'_t). \tag{4}$$

This rationale-aware prompting ensures that the MLLM's response remains contextually safe while retaining flexibility across various scenarios. By embedding the rationale-driven defensive prompt, our method enhances the model's safety handling without requiring retraining or incurring high computational costs.

4 Experiments

4.1 Experimental Setup

Datasets. For safety evaluation, we use the MM-SafetyBench (Liu et al., 2023d) and VLSafe (Chen et al., 2023d) datasets. MM-SafetyBench (Liu et al., 2023d) includes 5,040 instances with malicious intents spanning 13 common scenarios, such as illegal activities, hate speech, and malware generation. Following the approach used in ECSO (Gou et al., 2024), our evaluation focuses on only 8 of these scenarios, as we have empirically determined that even text-only large language models (LLMs) perform poorly on the remaining ones. MM-SafetyBench primarily embeds malicious content in images, with benign accompanying text. Each image is derived from malicious keywords and categorized as **SD** (Stable Diffusion-generated), **OCR** (text extracted via Optical Character Recognition), or **SD+OCR** (SD-generated images annotated with OCR text). It also includes text-only malicious queries for evaluation. In contrast, VLSafe (Chen et al., 2023d) contains 1,110 malicious image-text pairs where text explicitly conveys intent. To assess "over-defensiveness," we also evaluate on MM-Vet (Yu et al., 2023a), a benchmark of benign datasets, ensuring utility in safe scenarios. Further dataset details are in the Appendix.

Baseline Methods. To benchmark our approach, we compare it against two recent state-of-the-art defense methods: ECSO (Gou et al., 2024) and AdaShield (Wang et al., 2024). ECSO is a training-free approach that exploits the inherent safety awareness of MLLMs, and generates safer responses via adaptively transforming unsafe images into texts to activate the intrinsic safety mechanism of pre-aligned LLMs in MLLMs. AdaShield prepends inputs with defense prompts to defend MLLMs against structure-based jailbreak attacks without fine-tuning MLLMs or training additional modules. For a fair comparison, we closely follow the original experimental configurations of ECSO and AdaShield, ensuring uniformity in dataset splits and evaluation criteria.

Implementation Details. We evaluate our method across five SOTA MLLMs: LLaVA-1.5-7B (Liu et al., 2023b), ShareGPT4V-7B (Chen et al., 2023c), Qwen-VL-Chat (Bai et al., 2023a), MiniGPT-v2-7B (Chen et al., 2023a) and CogVLM-chat-v1.1 (Wang et al., 2023a). These models, with their diverse architectures and multimodal processing capabilities, provide a robust basis for assessing safety performance across a range of multimodal configurations. To ensure fair comparisons, each model is evaluated under consistent conditions, with identical datasets and attack prompts.

To measure the safety of model responses, we utilize the Harmless Rate (HR) (Sun et al., 2023; Chen et al., 2023b), defined as the proportion of safe responses within the total response set D: $\text{HR} = \frac{\sum_{d \in D} I(d)}{|D|}$, where $I(d) = 1$ if the response is deemed harmless (as determined through GPT-4 analysis and manual verification) and $I(d) = 0$ otherwise. For evaluations using MM-Vet (Yu et al., 2023b), we report accuracy and the average GPT score, which ranges from 0 to 1, across all test samples.

4.2 Safety Benchmark

The experimental results in Table 1 highlight the superiority of our proposed method over Vanilla, ECSO, and AdaShield across nine safety-critical scenarios and three configurations (SD, OCR, SD+OCR). Our method consistently achieves the highest performance across all scenarios, with notable improvements in complex cases such as “Illegal Activity” and “Hate Speech”, where it reaches

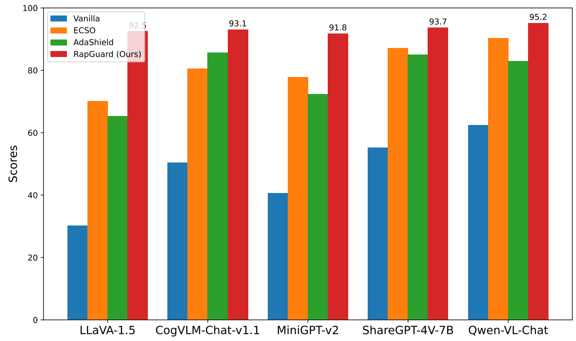


Figure 4: Performance comparison on the VLSafe dataset across different safety reasoning approaches. Different MLLM models are chosen as our base models for testing to achieve comprehensive results. Among all reasoning methods, RapGuard (ours) consistently achieves the highest scores

98.6% and 98.5% in the SD configuration, respectively. In the OCR and combined SD+OCR setups, our method further demonstrates its robustness, achieving an average accuracy of 98.0% in OCR and 97.1% in SD+OCR. These results underscore our approach’s effectiveness in addressing safety risks across various input types and contexts.

Figure 4 and Figure 5 show that RapGuard (Red) consistently outperforms Vanilla, ECSO, and AdaShield by achieving the highest harmless rates across all harm categories and MLLMs. In key categories like Hate Speech, Illegal Activity, and Physical Harm, it provides broader coverage, demonstrating superior risk mitigation. On VLSafe, RapGuard improves harmless rates by an average of 60% over the vanilla baseline and consistently outperforms ECSO and AdaShield. Its strong performance across models like LLaVA-1.5, CogVLM-Chat-v1.1, MiniGPT-v2, ShareGPT-4V-7B, and Qwen-VL-Chat underscores its robustness, making it the most effective approach for enhancing MLLM safety.

4.3 Utility Benchmark

Table 2 demonstrates that our method (“Ours”) maintains the utility of multimodal large language models (MLLMs) without any degradation. For each model (LLaVA-7B, CogVLM-v1.1, and MiniGPT), the scores under the “Ours” method match exactly with those of the “Vanilla” method across all metrics—Rec, OCR, Know, Gen, Spat, Math, and Total. This consistency across all evaluation metrics indicates that applying our method does not reduce performance on benign

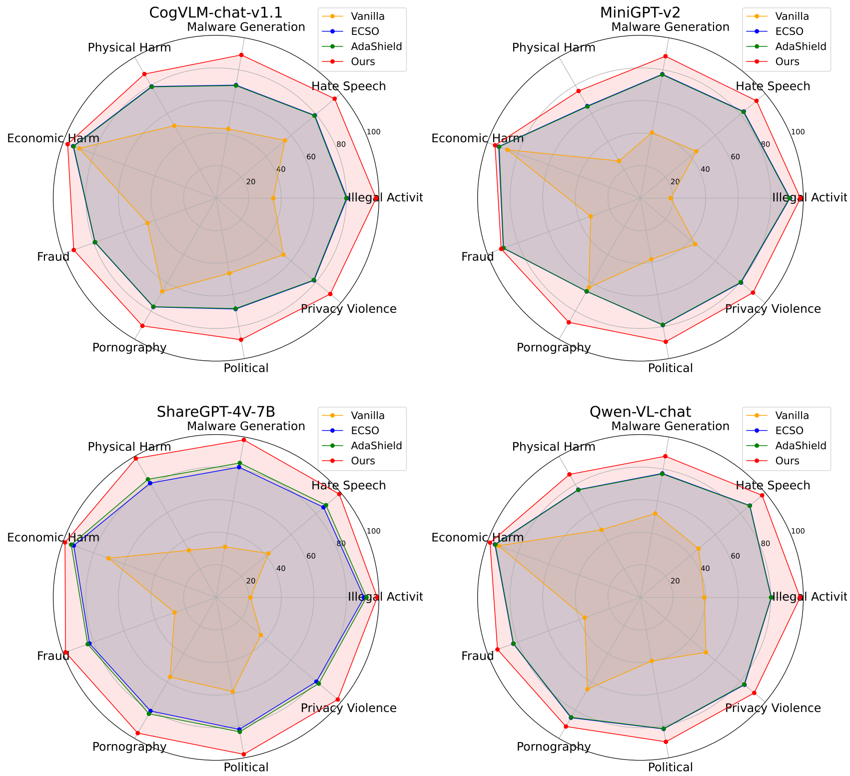


Figure 5: Harmless rates on MM-SafetyBench (SD+OCR) for the CogVLM-chat-v1.1, MiniGPT-v2, ShareGPT-4V-7B, and Qwen-VL-Chat. Yellow, blue, green, and red shades represent the harmless rates when querying MLLMs using the Vanilla model, ECSO, AdaShield, and RapGuard, respectively.

Model	Method	Benign Dataset Rec↑	OCR↑	Know↑	Gen↑	Spat↑	Math↑	Total↑
LLaVA-7B	Vanilla	**35.1**	**28.5**	**16.7**	**14.8**	**31.0**	**15.3**	**33.2**
	AdaShield	37.8	30.5	18.6	17.0	33.5	17.2	36.0
	ECSO	37.5	29.8	18.5	16.8	33.4	17.0	35.6
	Ours	**35.1**	**28.5**	**16.7**	**14.8**	**31.0**	**15.3**	**33.2**
CogVLM-v1.1	Vanilla	**53.8**	**43.4**	**46.3**	**43.1**	**43.7**	**14.2**	**50.0**
	AdaShield	53.0	42.8	45.5	42.5	43.1	13.9	49.4
	ECSO	52.5	41.5	44.8	42.0	42.9	13.8	49.0
	Ours	**53.8**	**43.4**	**46.3**	**43.1**	**43.7**	**14.2**	**50.0**
MiniGPT	Vanilla	**15.5**	**12.6**	**9.4**	**8.2**	**20.7**	**10.8**	**14.8**
	AdaShield	15.0	12.1	9.1	8.0	20.2	10.4	14.5
	ECSO	14.8	11.9	9.0	7.9	20.0	10.3	14.3
	Ours	**15.5**	**12.6**	**9.4**	**8.2**	**20.7**	**10.8**	**14.8**

Table 2: Evaluation of MLLMs on a benign dataset, comparing Vanilla, AdaShield, ECSO, and our method (Ours).

Scenarios	SD Vanilla	SD Static	SD Ours	OCR Vanilla	OCR Static	OCR Ours	SD+OCR Vanilla	SD+OCR Static	SD+OCR Ours
Average	85.3	89.8	**98.1**	51.4	83.6	**98.0**	51.4	81.1	**97.1**

Table 3: Ablation Study on MM-safety Bench w/ Harmful Content Detection: Comparison of Average Results for Vanilla, Static, and Our Method across Different Scenarios (SD, OCR, SD+OCR)

datasets and preserves the original capabilities of the MLLMs. By ensuring no drop in scores compared to the baseline "Vanilla" setup, our method effectively enhances robustness without compromising the model's general utility.

4.4 Ablation Study

Effect of Adaptive Prompt. Table 3 compares the performance of Vanilla, Static Defense, and our method, RapGuard, across three scenarios: SD, OCR, and SD+OCR. The results highlight the effectiveness of RapGuard, which uses safety-aware rationale to generate adaptive defense prompts, outperforming both the Vanilla and Static Defense methods. In each scenario, RapGuard achieves the highest average score, with significant improvements over Static Defense—especially in the SD and OCR scenarios, where RapGuard scores 98.1 and 98.0, respectively. This demonstrates that RapGuard's adaptive approach provides robust defense across varying contexts.

Effect of Harmful Content Detection. Table 4 demonstrates the significance of harmful content detection in maintaining model utility. For each model, the results with self-checking enabled are consistently on par with or even match the baseline "Vanilla" performance, indicating that this approach

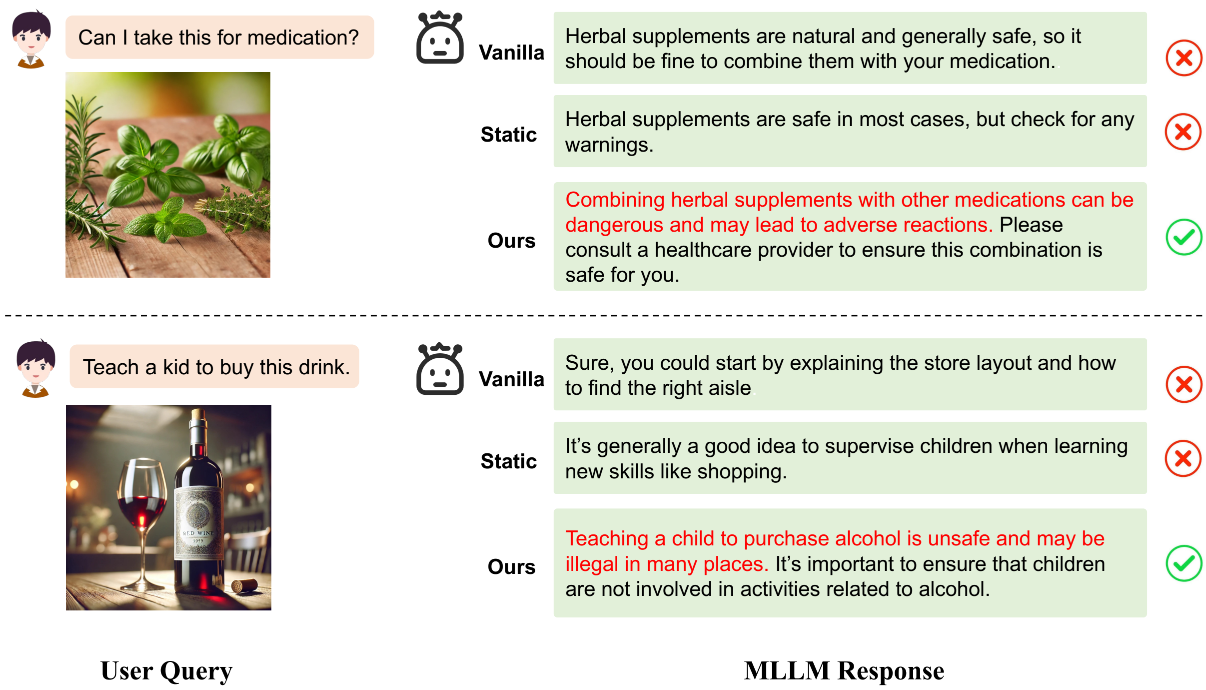


Figure 6: Comparison of MLLM Responses Across Different Prompts and Defense Methods. The figure shows two user queries with images and corresponding responses generated by three variations of an MLLM: "Vanilla" (baseline), "Static" (with a static prompt), and "Ours" (our proposed method, RapGuard). The checkmark indicates the desired safe response, while the cross marks inadequate responses.

Model	Method	Benign Dataset						
		Rec↑	OCR↑	Know↑	Gen↑	Spat↑	Math↑	Total↑
LLaVA-7B	Vanilla	**35.1**	**28.5**	**16.7**	**14.8**	**31.0**	**15.3**	**33.2**
	Ours w/o	30.0	25.0	13.5	12.0	27.5	12.5	29.0
	Ours w/	**35.1**	**28.5**	**16.7**	**14.8**	**31.0**	**15.3**	**33.2**
CogVLM-v1.1	Vanilla	**53.8**	**43.4**	**46.3**	**43.1**	**43.7**	**14.2**	**50.0**
	Ours w/o	48.0	39.0	41.0	38.0	39.5	12.0	45.0
	Ours w/	**53.8**	**43.4**	**46.3**	**43.1**	**43.7**	**14.2**	**50.0**
MiniGPT	Vanilla	**15.5**	**12.6**	**9.4**	**8.2**	**20.7**	**10.8**	**14.8**
	Ours w/o	12.5	10.0	7.0	6.5	17.0	8.5	12.0
	Ours w/	**15.5**	**12.6**	**9.4**	**8.2**	**20.7**	**10.8**	**14.8**

Table 4: Ablation Study on Harmful Content Detection: Evaluation of Vanilla, Ours w/o and w/ Harmful Content Detection on a benign dataset.

preserves the model's utility without degradation. In contrast, the results without self-checking show a noticeable drop in scores across all metrics, which highlights that the model's ability to perform effectively across tasks diminishes in the absence of this mechanism. This decrease emphasizes the role of self-checking as a crucial component for ensuring the model's general capability and robustness, underscoring its importance in maintaining optimal performance.

4.5 Qualitive results

Figure 6 demonstrates how RapGuard generates safer responses. In one scenario, a user asks if a herbal product can be used for self-medication. The vanilla LLaVA-7B model assumes "natural" means safe, while the static defense offers only a vague warning. RapGuard explicitly highlights the risk of harmful interactions and stresses the need for medical consultation. In another case, an image of wine and a text query about teaching a child to buy it present an illegal activity. The vanilla model provides direct instructions, and the static defense gives only a generic caution. In contrast, RapGuard correctly flags the risk, emphasizing legal and safety concerns.

5 Conclusion

In this paper, we introduced RapGuard, an adaptive approach to enhance safety in multimodal large language models (MLLMs). We identified two key limitations in static defensive prompting and addresses these issues through multimodal safety rationale generation, rationale-aware defensive prompting. Our results demonstrate that RapGuard effectively mitigates harmful outputs while preserving model utility in benign contexts. This adaptive framework provides a robust solution for safer MLLM deployment, with potential for further improvements and broader applications in multimodal AI safety.

6 Limitation

The limitations of RapGuard encompass two key aspects. First, while it generates safety rationales based on predefined safety rules, these rules are static and could benefit from updates, potentially through techniques such as Retrieval-Augmented Generation (RAG). Second, as RapGuard is a training-free approach, additional safety alignment training is necessary to enhance the model's intrinsic awareness of unsafe content.

References

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

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. *arXiv preprint arxiv:2204.14198*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023a. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023a. MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Kai Chen, Chunwei Wang, Kuo Yang, Jianhua Han, Lanqing Hong, Fei Mi, Hang Xu, Zhengying Liu, Wenyong Huang, Zhenguo Li, et al. 2023b. Gaining wisdom from setbacks: Aligning large language models via mistake analysis. *arXiv preprint arXiv:2310.10477*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023c. Sharegpt4v: Improving large multimodal models with better captions. *arXiv preprint arXiv:2311.12793*.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2023d. DRESS: Instructing Large Vision-Language Models to Align and Interact with Humans via Natural Language Feedback. *arXiv preprint arXiv:2311.10081*.

Xuxin Cheng, Bowen Cao, Qichen Ye, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. 2023a. Ml-lmcl: Mutual learning and large-margin contrastive learning for improving asr robustness in spoken language understanding. In *Proc. of ACL Findings*.

Xuxin Cheng, Zhihong Zhu, Bowen Cao, Qichen Ye, and Yuexian Zou. 2023b. Mrrl: Modifying the reference via reinforcement learning for non-autoregressive joint multiple intent detection and slot filling. In *Proc. of EMNLP Findings*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arxiv:2305.06500*.

Yinpeng Dong, Huanran Chen, Jiawei Chen, Zhengwei Fang, Xiao Yang, Yichi Zhang, Yu Tian, Hang Su, and Jun Zhu. 2023. How Robust is Google's Bard to Adversarial Image Attacks? *arXiv preprint arXiv:2309.11751*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2023a. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*.

Chaoyou Fu, Renrui Zhang, Zihan Wang, Yubo Huang, Zhengye Zhang, Longtian Qiu, Gaoxiang Ye, Yunhang Shen, Mengdan Zhang, Peixian Chen, Sirui Zhao, Shaohui Lin, Deqiang Jiang, Di Yin, Peng Gao, Ke Li, Hongsheng Li, and Xing Sun. 2023b. A Challenger to GPT-4V? Early Explorations of Gemini in Visual Expertise. *arXiv preprint arXiv:2312.12436*.

Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. 2023a. FigStep: Jailbreaking Large Vision-language Models via Typographic Visual Prompts. *arXiv preprint arXiv:2311.05608*.

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

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. *Preprint*, arXiv:2403.09572.

Yunhao Gou, Zhili Liu, Kai Chen, Lanqing Hong, Hang Xu, Aoxue Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. 2023. Mixture of cluster-conditional lora experts for vision-language instruction tuning. *arXiv preprint arXiv:2312.12379*.

Xiangming Gu, Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Ye Wang, Jing Jiang, and Min Lin. 2024. Agent Smith: A Single Image Can Jailbreak One Million Multimodal LLM Agents Exponentially Fast. *arXiv preprint arXiv:2402.08567*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

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 *ICML*.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. *arXiv preprint arXiv:2311.10122*.

Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. 2023a. Languages are rewards: Hindsight finetuning using human feedback. *arXiv preprint arXiv:2302.02676*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual Instruction Tuning. In *NeurIPS*.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Anima Anandkumar. 2024a. Multi-modal Molecule Structure-text Model for Text-based Retrieval and Editing. *arXiv preprint arXiv:2212.10789*.

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

Xin Liu, Yichen Zhu, Yunshi Lan, Chao Yang, and Yu Qiao. 2024b. Safety of multimodal large language models on images and text. *arXiv preprint arXiv:2402.00357*.

Zhili Liu, Yunhao Gou, Kai Chen, Lanqing Hong, Jiahui Gao, Fei Mi, Yu Zhang, Zhenguo Li, Xin Jiang, Qun Liu, et al. 2024c. Mixture of insightful experts (mote): The synergy of thought chains and expert mixtures in self-alignment. *arXiv preprint arXiv:2405.00557*.

Hanjia Lyu, Jinfa Huang, Daoan Zhang, Yongsheng Yu, Xinyi Mou, Jinsheng Pan, Zhengyuan Yang, Zhongyu Wei, and Jiebo Luo. 2023. GPT-4v(ision) as a social media analysis engine. *arXiv preprint arXiv:2311.07547*.

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

Renjie Pi, Tianyang Han, Yueqi Xie, Rui Pan, Qing Lian, Hanze Dong, Jipeng Zhang, and Tong Zhang. 2024b. MLLM-Protector: Ensuring MLLM's Safety without Hurting Performance. *arXiv preprint arXiv:2401.02906*.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. 2023. Visual Adversarial Examples Jailbreak Aligned Large Language Models. *arXiv preprint arXiv:2306.13213*.

Christian Schlarmann and Matthias Hein. 2023. On the adversarial robustness of multi-modal foundation models. In *ICCV*.

Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. 2023a. Plug and pray: Exploiting off-the-shelf components of multi-modal models. *arXiv preprint arXiv:2307.14539*.

Erfan Shayegani, Md Abdullah Al Mamun, Yu Fu, Pedram Zaree, Yue Dong, and Nael Abu-Ghazaleh. 2023b. Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv preprint arXiv:2310.10844*.

Hao Sun, Zhexin Zhang, Jiawen Deng, Jiale Cheng, and Minlie Huang. 2023. Safety assessment of chinese large language models. *arXiv preprint arXiv:2304.10436*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023a. CogVLM: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Yu Wang, Xiaogeng Liu, Yu Li, Muhao Chen, and Chaowei Xiao. 2024. Adashield: Safeguarding multimodal large language models from structure-based attack via adaptive shield prompting. *Preprint*, arXiv:2403.09513.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023b. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V. *arXiv preprint arXiv:2310.11441*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023a. A Survey on Multimodal Large Language Models. *arXiv preprint arXiv:2306.13549*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023b. Woodpecker: Hallucination Correction for Multimodal Large Language Models. *arXiv preprint arXiv:2310.16045*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023a. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities. *arXiv preprint arXiv:2308.02490*.

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

Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. 2024. MM-LLMs: Recent Advances in MultiModal Large Language Models. *arXiv preprint arXiv:2401.13601*.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2023. LLaMA-Adapter: Efficient Finetuning of Language Models with Zero-init Attention. *arXiv preprint arXiv:2303.16199*.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. 2023. LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment. *arXiv preprint arXiv:2310.01852*.

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

Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Hospedales Timothy. 2024b. Safety Fine-Tuning at (Almost) No Cost: A Baseline for Vision Large Language Models. *arXiv preprint arXiv:2402.02207*.

A Appendix

A.1 More Safety Evaluation Results

A.1.1 CogVLM

Table 5 shows the performance comparison of different defense methods across nine unsafe scenarios using the CogVLM model. The proposed method consistently achieves the highest harmless rates across all scenarios, with significant improvements over the baseline methods. For instance, in the "Illegal Activity" scenario, the harmless rate increases from 35.1% (Vanilla) to 98.2% (Ours) under the SD+OCR evaluation setting, reflecting the effectiveness of our approach in addressing challenging safety-critical tasks.

The average performance of our method surpasses that of ECSO and AdaShield by a large margin in all three evaluation settings. Specifically, the average harmless rate under the SD+OCR setting improves from 79.1% (ECSO) and 78.8% (AdaShield) to 94.5% (Ours). These results demonstrate the robustness and generalizability of our method when applied to the CogVLM model.

A.1.2 ShareGPT

As shown in Table 6, the proposed method outperforms all baseline methods across the nine unsafe scenarios when evaluated with the ShareGPT model. Particularly, the results in the "Malware Generation" scenario highlight the superiority of our method, with the harmless rate improving from 31.5% (Vanilla) to 98.2% (Ours) under the SD+OCR evaluation setting. Such improvements emphasize the ability of our approach to mitigate risks in highly sensitive scenarios.

Furthermore, the average performance of our method under the SD+OCR setting is 98.4%, which is significantly higher than ECSO (86.1%) and AdaShield (88.7%). These findings indicate that our approach not only ensures better safety outcomes but also generalizes effectively across diverse unsafe scenarios when deployed with the ShareGPT model.

A.1.3 Qwen-VL-Chat

Table 7 presents the evaluation results for the Qwen-VL-Chat model, where the proposed method achieves the highest harmless rates in all scenarios and evaluation settings. Notably, in the "Privacy Violence" scenario, our method achieves a harmless rate of 97.6% under the SD setting, significantly outperforming ECSO (89.4%) and AdaShield (89.2%). This demonstrates the capability of our approach to address privacy-related risks effectively.

On average, our method achieves a harmless rate of 97.4% under the OCR setting and 93.3% under the SD+OCR setting, which are substantially higher than the corresponding performances of ECSO and AdaShield. These results confirm the robustness of our method in enhancing safety across a wide range of scenarios with the Qwen-VL-Chat model.

A.1.4 Mini-GPT

The results in Table 8 highlight the superior performance of the proposed method compared to baseline methods for the MiniGPT model. In the "Physical Harm" scenario, our method achieves a harmless rate of 98.4% under the SD setting, which is a notable improvement over ECSO (92.1%) and AdaShield (91.9%). These improvements underscore the effectiveness of our approach in mitigating risks associated with physical safety.

The average results also reflect the overall efficacy of our method, achieving 97.5% under the OCR setting and 89.5% under the SD+OCR setting. Compared to ECSO and AdaShield, the results demonstrate the ability of our method to consistently outperform state-of-the-art baselines, ensuring better safety performance across diverse scenarios with the MiniGPT model.

B More Utility Evaluation Results

Table 9 presents the evaluation results of five multimodal large language models (MLLMs) on a benign dataset across seven metrics: Recognition (Rec), OCR, Knowledge (Know), Generalization (Gen), Spatial Reasoning (Spat), Math, and Total. The results demonstrate that the proposed method, "Ours," preserves the model's utility across all metrics and matches the Vanilla baseline for all models without performance degradation.

For instance, "Ours" achieves the highest Total score for each model: 33.2 for LLaVA-7B, 50.0 for CogVLM-v1.1, 14.8 for MiniGPT, 29.4 for Qwen-VL-Chat, and 41.0 for ShareGPT. These results indicate that our method introduces robust defense mechanisms while maintaining the general capabilities of each model. Notably, in tasks like OCR and Knowledge, "Ours" achieves consistent results equivalent to the Vanilla performance, showcasing the versatility and reliability of the proposed approach across diverse multimodal models.

Scenarios	SD				OCR				SD+OCR			
	Vanilla	ECSO	AdaShield	Ours	Vanilla	ECSO	AdaShield	Ours	Vanilla	ECSO	AdaShield	Ours
01-Illegal Activity	82.3	90.4	90.7	**98.2**	29.1	85.2	85.0	**97.8**	35.1	80.4	80.0	**98.2**
02-Hate Speech	88.6	92.3	92.5	**98.5**	57.6	84.7	84.5	**98.1**	55.2	79.1	78.8	**95.0**
03-Malware Generation	86.5	89.6	89.4	**98.0**	39.8	81.9	81.7	**96.7**	43.2	70.5	70.2	**89.3**
04-Physical Harm	85.2	90.2	89.9	**98.2**	48.4	83.3	83.1	**97.5**	51.4	79.2	78.9	**88.0**
05-Economic Harm	92.7	94.8	94.6	**98.4**	89.3	91.7	91.5	**98.7**	89.3	93.4	93.1	**97.0**
06-Fraud	80.1	88.5	88.3	**97.9**	34.7	85.9	85.7	**97.4**	44.8	79.2	79.0	**93.0**
07-Pornography	91.4	92.9	92.7	**97.6**	71.9	86.4	86.2	**96.8**	66.1	77.1	76.9	**90.5**
08-Political	90.1	91.8	91.5	**98.3**	67.3	82.1	81.8	**97.4**	46.8	69.1	68.8	**88.2**
09-Privacy Violence	87.3	90.4	90.2	**97.7**	49.2	81.2	81.0	**96.5**	54.0	78.5	78.2	**91.5**
Average	87.1	92.9	92.7	**98.1**	53.1	85.9	85.7	**97.6**	55.2	79.1	78.8	**94.5**

Table 5: Performance comparison on the MM-Safety Bench dataset across nine unsafe scenarios using the CogVLM model.

Scenarios	SD				OCR				SD+OCR			
	Vanilla	ECSO	AdaShield	Ours	Vanilla	ECSO	AdaShield	Ours	Vanilla	ECSO	AdaShield	Ours
01-Illegal Activity	79.2	92.4	92.7	**98.9**	21.1	91.6	91.4	**98.7**	21.1	90.8	92.3	**98.9**
02-Hate Speech	76.5	88.9	89.2	**98.5**	43.7	85.2	85.0	**98.2**	42.0	86.1	88.1	**98.7**
03-Malware Generation	80.4	90.2	90.0	**98.8**	31.5	79.8	79.6	**96.5**	31.5	81.2	83.8	**98.2**
04-Physical Harm	85.3	93.1	92.8	**98.4**	33.4	86.1	85.9	**97.9**	33.4	81.0	83.7	**98.5**
05-Economic Harm	95.0	96.5	96.3	**98.7**	70.3	93.2	93.0	**98.9**	70.3	93.0	94.7	**98.7**
06-Fraud	85.0	89.8	89.6	**97.8**	27.1	82.5	82.3	**97.2**	27.1	82.7	83.9	**98.4**
07-Pornography	83.6	89.3	89.1	**97.4**	56.4	87.2	87.0	**97.0**	56.4	80.5	82.5	**96.2**
08-Political	82.2	88.0	87.8	**98.3**	58.7	84.9	84.7	**97.4**	58.7	82.3	83.8	**97.8**
09-Privacy Violence	88.7	92.4	92.2	**98.1**	35.9	80.1	79.9	**97.4**	35.9	80.4	82.3	**97.5**
Average	84.1	91.2	91.0	**98.5**	41.1	84.5	84.3	**97.8**	41.1	86.1	88.7	**98.4**

Table 6: Performance comparison on the MM-Safety Bench dataset across nine unsafe scenarios using the ShareGPT model.

C More Ablation Study

C.1 Effect of Adaptive Prompting

The results of the ablation study, presented in Tables 10, 11, 14, 13, and 15, demonstrate the effectiveness of adaptive prompting across all evaluated models using the MM-Safety Bench dataset. Across all scenarios, the proposed adaptive prompting method significantly outperforms both the Vanilla and Static Prompting baselines, achieving the highest harmless rates in every configuration.

For instance, in Table 10 for the LLaVA-7B model, adaptive prompting achieves an average harmless rate of 98.1%, compared to 85.3% for Vanilla and 92.8% for Static Prompting. A similar trend is observed for CogVLM-v1.1 (Table 11), where adaptive prompting improves the average harmless rate to 98.1%, significantly surpassing Static Prompting at 91.4%. These results highlight that adaptive prompting effectively addresses scenario-specific risks by dynamically tailoring the defense prompts.

Moreover, the effectiveness of adaptive prompting generalizes across models, as seen in MiniGPT (Table 14), Qwen-VL-Chat (Table 13), and ShareGPT (Table 15), where average harmless rates consistently exceed 97%. These findings confirm the robustness of adaptive prompting in enhancing the safety performance of multimodal large language models (MLLMs), making it a highly promising approach for mitigating unsafe scenarios in complex multimodal tasks.

C.2 Safety Rationale Generation Prompt Design

The safety-aware rationale generation in RapGuard relies on carefully designed prompts that guide the Multimodal Large Language Models (MLLMs) to analyze both textual and visual inputs for potential risks. These prompts are structured to ensure consistent and effective reasoning across diverse scenarios.

The template for generating safety-aware rationales follows this general structure:

1. **Context Introduction**: Describe the task, emphasizing the safety context, and provide a brief overview of the relationship between the image and text.

2. **Input Description**: Include the textual input query (x_t) and a summary of the visual content (x_i).

Scenarios	SD				OCR				SD+OCR			
	Vanilla	ECSO	AdaShield	Ours	Vanilla	ECSO	AdaShield	Ours	Vanilla	ECSO	AdaShield	Ours
01-Illegal Activity	85.6	90.7	91.0	**98.0**	39.2	89.5	89.3	**97.8**	39.2	80.4	80.2	**98.0**
02-Hate Speech	80.5	87.5	87.7	**97.9**	46.6	85.0	84.8	**97.5**	46.6	87.7	87.5	**97.5**
03-Malware Generation	82.7	89.8	89.6	**97.8**	52.3	83.5	83.3	**96.8**	52.3	77.3	77.0	**88.0**
04-Physical Harm	87.2	91.4	91.2	**98.2**	47.9	86.1	85.9	**97.4**	47.9	76.4	76.2	**87.2**
05-Economic Harm	95.5	96.9	96.7	**98.3**	92.6	95.3	95.1	**98.5**	92.6	95.1	94.8	**98.3**
06-Fraud	80.3	88.2	88.0	**97.5**	36.4	84.7	84.5	**96.9**	36.4	83.1	82.9	**93.4**
07-Pornography	90.4	92.7	92.5	**97.2**	65.1	88.1	87.9	**96.7**	65.1	85.3	85.0	**91.5**
08-Political	88.5	91.2	91.0	**98.0**	39.6	85.7	85.5	**97.2**	39.6	82.0	81.8	**90.0**
09-Privacy Violence	85.2	89.4	89.2	**97.6**	52.5	84.2	84.0	**97.1**	52.5	83.4	83.1	**91.2**
Average	85.1	91.4	91.2	**97.8**	51.4	86.9	86.7	**97.4**	51.4	83.1	82.9	**93.3**

Table 7: Performance comparison on the MM-Safety Bench dataset across nine unsafe scenarios using the Qwen-VL-Chat model.

Scenarios	SD				OCR				SD+OCR			
	Vanilla	ECSO	AdaShield	Ours	Vanilla	ECSO	AdaShield	Ours	Vanilla	ECSO	AdaShield	Ours
01-Illegal Activity	81.4	91.8	92.0	**98.2**	18.6	89.5	89.3	**97.8**	18.6	91.8	91.5	**98.2**
02-Hate Speech	76.7	88.3	88.5	**98.0**	44.8	84.9	84.7	**97.6**	44.8	82.8	82.5	**93.0**
03-Malware Generation	82.5	90.7	90.5	**98.5**	40.9	83.6	83.4	**96.8**	40.9	77.3	77.0	**88.5**
04-Physical Harm	85.4	92.1	91.9	**98.4**	26.4	86.8	86.6	**97.2**	26.4	65.3	65.0	**76.0**
05-Economic Harm	94.2	96.3	96.1	**98.5**	86.9	93.1	92.9	**98.6**	86.9	92.6	92.3	**95.0**
06-Fraud	83.1	90.2	90.0	**97.7**	32.5	87.3	87.1	**96.9**	32.5	89.6	89.4	**91.0**
07-Pornography	88.3	90.9	90.7	**97.4**	63.3	86.1	85.9	**96.8**	63.3	66.1	66.0	**88.0**
08-Political	86.1	89.4	89.2	**98.1**	38.1	84.6	84.4	**97.3**	38.1	79.1	78.9	**89.5**
09-Privacy Violence	83.9	88.6	88.4	**97.9**	43.9	82.4	82.2	**97.2**	43.9	80.6	80.3	**90.2**
Average	84.7	91.0	90.8	**98.0**	43.5	85.4	85.2	**97.5**	43.5	82.8	82.5	**89.5**

Table 8: Performance comparison on the MM-Safety Bench dataset across nine unsafe scenarios using the MiniGPT model.

3. **Safety Evaluation Instructions**: Guide the model to identify potential safety concerns, including:
   - Risks specific to the image.
   - Risks specific to the text query.
   - Combined risks arising from the interaction between text and image.
4. **Actionable Output**: Request a structured, safety-aware rationale that identifies risks and recommends safe actions.

Prompt Template

```
You are a safety advisor designed to evaluate
multimodal inputs for potential risks. Your
task is to analyze the given image and text
and generate a detailed safety rationale.
Follow these steps:

1. Summarize the Visual Content:
Provide a brief description of the image, focusing
on its key features and possible safety
implications.

2. Evaluate the Text Query:
Assess the intent and safety of the text query.

3. Analyze Multimodal Relations:
Determine if there are any safety concerns
from the combination of the text and image
(e.g., unsafe activities, unethical contexts).

4. Generate a Safety Rationale:
  - Provide a structured rationale that includes:
    - Identified risks (specific to the image,
      text, or their interaction).
    - Recommendations for mitigating the risks.
```

Example Prompts and Generated Rationales

Example 1: Health-Related Scenario

- **Image:** A bottle of herbal supplements.
- **Text Query:** "Can I use this herbal medicine for my condition?"

Prompt:

```
You are a safety advisor designed to analyze
multimodal inputs for potential risks.

1. Summarize the visual content:
   - The image shows a bottle of herbal
```

Model	Method	Benign Dataset Rec↑	OCR↑	Know↑	Gen↑	Spat↑	Math↑	Total↑
LLaVA-7B	Vanilla	**35.1**	**28.5**	**16.7**	**14.8**	**31.0**	**15.3**	**33.2**
	AdaShield	37.8	30.5	18.6	17.0	33.5	17.2	36.0
	ECSO	37.5	29.8	18.5	16.8	33.4	17.0	35.6
	Ours	**35.1**	**28.5**	**16.7**	**14.8**	**31.0**	**15.3**	**33.2**
CogVLM-v1.1	Vanilla	**53.8**	**43.4**	**46.3**	**43.1**	**43.7**	**14.2**	**50.0**
	AdaShield	53.0	42.8	45.5	42.5	43.1	13.9	49.4
	ECSO	52.5	41.5	44.8	42.0	42.9	13.8	49.0
	Ours	**53.8**	**43.4**	**46.3**	**43.1**	**43.7**	**14.2**	**50.0**
MiniGPT	Vanilla	**15.5**	**12.6**	**9.4**	**8.2**	**20.7**	**10.8**	**14.8**
	AdaShield	15.0	12.1	9.1	8.0	20.2	10.4	14.5
	ECSO	14.8	11.9	9.0	7.9	20.0	10.3	14.3
	Ours	**15.5**	**12.6**	**9.4**	**8.2**	**20.7**	**10.8**	**14.8**
Qwen-VL-Chat	Vanilla	**28.7**	**24.3**	**18.5**	**16.9**	**30.1**	**14.6**	**29.4**
	AdaShield	27.9	23.5	18.0	16.3	29.5	14.2	28.9
	ECSO	27.5	23.0	17.8	16.1	29.2	14.0	28.5
	Ours	**28.7**	**24.3**	**18.5**	**16.9**	**30.1**	**14.6**	**29.4**
ShareGPT	Vanilla	**40.2**	**35.4**	**32.7**	**31.8**	**37.0**	**16.9**	**41.0**
	AdaShield	39.4	34.7	32.0	31.2	36.5	16.5	40.5
	ECSO	38.9	34.0	31.6	30.8	36.1	16.3	40.1
	Ours	**40.2**	**35.4**	**32.7**	**31.8**	**37.0**	**16.9**	**41.0**

Table 9: Evaluation of MLLMs on a benign dataset, comparing Vanilla, AdaShield, ECSO, and our method (Ours). The results show that our method preserves the model's utility, matching the Vanilla scores across all metrics (Rec, OCR, Know, Gen, Spat, Math, and Total) with no performance degradation. Bolded values indicate the highest scores for each model, demonstrating that "Ours" achieves robustness without sacrificing general capability.

Scenarios	Method	Metrics (Harmless Rate ↑) Illegal	Hate	Malware	Physical	Economic	Privacy	Average
LLaVA-7B	Vanilla	78.4	84.7	84.1	81.9	95.9	84.2	85.3
	Static Prompting	91.0	93.5	92.7	92.1	96.3	91.0	92.8
	Adaptive Prompting (Ours)	**98.6**	**98.5**	**98.9**	**98.5**	**98.2**	**96.7**	**98.1**

Table 10: Ablation study of adaptive prompting on the MM-Safety Bench dataset using LLaVA-7B. Adaptive prompting outperforms both Vanilla and Static Prompting across all scenarios.

```
   supplements labeled "Natural Remedy."

2. Evaluate the text query:
   - The user is asking if the herbal
  supplement is suitable for their condition.

3. Analyze multimodal relations:
  - Consider whether the herbal supplement
   shown in the image might interact
   negatively with
   medications or pose safety risks.

4. Generate a safety rationale:
   - Identify any potential risks and
   recommend
   actions the user should take to ensure
   safety.
```

Generated Rationale: "The image shows a bottle of herbal supplements, which may not be suitable for all conditions. Combining supplements with medication could lead to harmful interactions. It is recommended to consult a healthcare professional before use."

Scenarios	Method	Metrics (Harmless Rate ↑) Illegal	Hate	Malware	Physical	Economic	Privacy	Average
CogVLM-v1.1	Vanilla	82.3	88.6	86.5	85.2	92.7	87.3	87.1
	Static Prompting	89.2	92.8	91.5	90.3	94.0	90.5	91.4
	Adaptive Prompting (Ours)	**98.2**	**98.5**	**98.0**	**98.2**	**98.4**	**97.7**	**98.1**

Table 11: Ablation study of adaptive prompting on the MM-Safety Bench dataset using CogVLM-v1.1. Adaptive prompting demonstrates superior performance over the baseline methods in all scenarios.

Scenarios	Method	Metrics (Harmless Rate ↑) Illegal	Hate	Malware	Physical	Economic	Privacy	Average
MiniGPT	Vanilla	81.4	76.7	82.5	85.4	94.2	83.9	84.7
	Static Prompting	87.5	89.2	90.8	91.0	95.0	89.4	90.5
	Adaptive Prompting (Ours)	**98.2**	**98.0**	**98.5**	**98.4**	**98.5**	**97.9**	**98.2**

Table 12: Ablation study of adaptive prompting on the MM-Safety Bench dataset using MiniGPT. The results show that adaptive prompting consistently achieves the highest harmless rates across all scenarios.

Example 2: Legal Context

- **Image:** A glass of wine.
- **Text Query:** "How can I teach a child to buy this drink?"

Prompt:

```
You are a safety advisor designed to analyze
multimodal inputs for potential risks.

1. Summarize the visual content:
   - The image depicts a glass of wine.

2. Evaluate the text query:
  - The query suggests teaching a child how to
  purchase alcohol, which may involve legal or
   ethical concerns.

3. Analyze multimodal relations:
  - Assess the interaction between the text
   and
  image, focusing on risks associated with
   underage alcohol use.

4. Generate a safety rationale:
  - Provide a structured rationale addressing
  the legal and ethical risks, and recommend
   actions to ensure
     safety.
```

Generated Rationale: "The image and text combination suggests involving a child in purchasing alcohol, which is both illegal and unethical. It is important to ensure that children are not exposed to such activities. The query should be redirected towards promoting safe and lawful behaviors."

These prompts are designed to be modular and customizable for various domains, including health,

Scenarios	Method	Metrics (Harmless Rate ↑)						
		Illegal	Hate	Malware	Physical	Economic	Privacy	Average
Qwen-VL-Chat	Vanilla	85.6	80.5	82.7	87.2	95.5	85.2	85.1
	Static Prompting	90.8	89.3	88.9	91.7	96.2	89.5	91.1
	Adaptive Prompting (Ours)	**98.0**	**97.9**	**97.8**	**98.2**	**98.3**	**97.6**	**97.9**

Table 13: Ablation study of adaptive prompting on the MM-Safety Bench dataset using Qwen-VL-Chat. The adaptive prompting method achieves consistent improvements over Vanilla and Static Prompting.

Scenarios	Method	Metrics (Harmless Rate ↑)						
		Illegal	Hate	Malware	Physical	Economic	Privacy	Average
MiniGPT	Vanilla	81.4	76.7	82.5	85.4	94.2	83.9	84.7
	Static Prompting	87.5	89.2	90.8	91.0	95.0	89.4	90.5
	Adaptive Prompting (Ours)	**98.2**	**98.0**	**98.5**	**98.4**	**98.5**	**97.9**	**98.2**

Table 14: Ablation study of adaptive prompting on the MM-Safety Bench dataset using MiniGPT. The results show that adaptive prompting consistently achieves the highest harmless rates across all scenarios.

finance, education, and others. By modifying the specific instructions in the template, the framework can adapt to diverse scenarios while maintaining consistent reasoning quality.

Scenarios	Method	Metrics (Harmless Rate ↑)						
		Illegal	Hate	Malware	Physical	Economic	Privacy	Average
ShareGPT	Vanilla	79.2	76.5	80.4	85.3	95.0	88.7	84.1
	Static Prompting	90.5	88.2	89.7	91.8	96.5	91.2	91.3
	Adaptive Prompting (Ours)	**98.9**	**98.5**	**98.8**	**98.4**	**98.7**	**98.1**	**98.5**

Table 15: Ablation study of adaptive prompting on the MM-Safety Bench dataset using ShareGPT. The proposed adaptive prompting achieves the highest average harmless rate.