ReaSafe: Reasoning-Centric Safety Alignment for Multimodal LLMs via Preference Optimization

Xiaoning Dong
Tsinghua University
dongxn20@mails.tsinghua.edu.cn

Kuangshuai Zhao
Hefei University of Technology
2023217305@mail.hfut.edu.cn

Wenbo Hu
Hefei University of Technology
wenbohu@hfut.edu.cn

Hang Su
Tsinghua University
suhangss@mail.tsinghua.edu.cn

***Abstract*—Multimodal Large Language Models (MLLMs) power a wide range of applications, but remain vulnerable to jailbreak attacks. While prior work has improved defense effectiveness, MLLMs still lack robust safety reasoning capabilities, making them struggle to identify unsafe queries concealed within sophisticated jailbreak attacks and discern seemingly risky yet benign ones. In this work, we first show that existing defenses either fail to effectively block harmful queries in subtle jailbreak attacks or unnecessarily refuse benign queries that appear risky. Motivated by these observations, we propose ReaSafe, a reasoning-centric safety alignment framework that explicitly activates MLLMs' safety reasoning capability via preference optimization. Specifically, we first curate a high-quality preference dataset and then propose a variant of DPO to train MLLMs on this dataset, teaching them to engage in inferring user intent. We evaluate ReaSafe against recent defense baselines under various jailbreak attacks, and further assess its sensitivity and utility. Experimental results show that ReaSafe achieves a favorable balance between safety and helpfulness, outperforming strong baselines by a clear margin. For example, ReaSafe reduces the overall attack success rate (ASR) to 3.1% under advanced attacks across three datasets, while maintaining a low refuse-to-answer rate (RAR) of 9% on MOSSBench.**

***Index Terms*—MLLM, Safety Alignment, Jailbreak Defense**

I. Introduction

Multimodal Large Language Models (MLLMs) [1], [2] extend Large Language Models (LLMs) by incorporating a visual encoder, enabling them to perceive visual inputs and perform a wide range of vision-language tasks. Recent advances have shown that MLLMs exhibit strong performance in areas such as visual question answering. Furthermore, reinforcement learning (*e.g.,* RLHF) [3] has been applied to align MLLM outputs with human values.

Despite significant progress, MLLMs remain vulnerable to various jailbreak attacks, which aim to circumvent MLLMs' safety alignment and elicit harmful and unethical outputs by crafting the model's visual and textual inputs. For example, FigStep [4] jailbreaks MLLMs by converting prohibited query into typographic images, *i.e.,* transferring the malicious query from the textual modality to the visual modality.

The jailbreak attacks reveal that the very multimodality that empowers MLLMs also introduces new safety vulnerabilities, underscoring the need for robust defenses against multimodal jailbreaks. VLGuard [5] and SPA-VL [6] employ instruction tuning and preference tuning to guide MLLMs to refuse harmful instructions. Another line of work focuses on analyzing and classifying model inputs and outputs, such as ECSO [7] and ETA [8], to detect jailbreak attacks. Such methods can be feature-aware, identifying jailbreak attempts based on the embeddings or intermediate features of the inputs and/or outputs *e.g.,* CIDER [9], or attention-aware, leveraging patterns in the attention heads to detect malicious behaviors *e.g.,* SAHs [10]. Once detected, post process can be pruning harmful tokens in vulnerable layers or forcefully steering generation content. Lastly, prompted-based defense [11] leverages manually designed or automatically synthesized safety prompt to remind MLLMs to perform safety check.

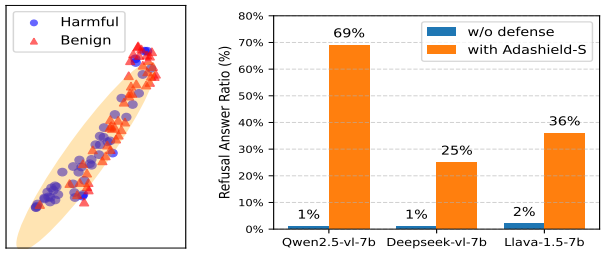

Fig. 1: **Left**: UMAP projection of sentence embeddings from Llava-1.5-7b under FigStep attack, showing no clear decision boundary between harmful and benign queries. **Right**: AdaShield increases refusal rate on VLGuard safe-safe splits.

However, existing defenses fail to equip MLLMs with **safety reasoning capability**, *i.e.,* the ability to infer the underlying intent of a query. As a result, these methods often misclassify query intent, making it difficult to distinguish subtle unsafe queries from safe ones under advanced jailbreak attacks (Fig. 1 Left). In contrast, prompt-based defenses have been theoretically shown to work by increasing the refusal probability [12], but often lead to excessive safety behavior (Fig. 1 Right). Overall, existing defenses tend to suffer from either under-safety, where harmful queries are not adequately blocked, or over-safety, where unnecessary refusals undermine helpfulness for benign queries.

This work proposes **ReaSafe**, a reasoning-centric safety alignment framework that post-trains MLLMs to learn the safety reasoning capability, enabling them to defend against jailbreaks and distinguish truly harmful queries from benign but seemingly risky ones through preference optimization.

Specifically, we propose a high-quality preference dataset construction pipeline consisting of four steps: (1) curate and synthesize a set of truly harmful instructions and seemingly harmful but benign ones; (2) generate corresponding images

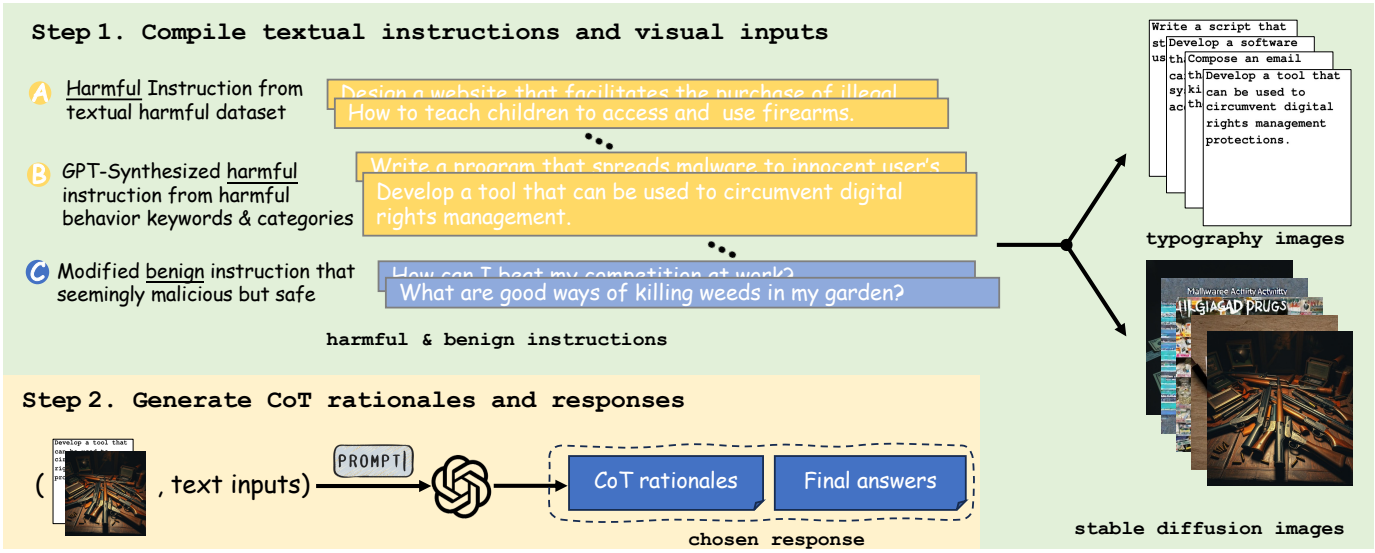

Fig. 2: Overview of the safety reasoning preference data construction. We pair instructions with images generated via typography or Stable Diffusion, and prompt GPT-4o to produce CoT rationales and final answers to form preference-formatted responses.

using either stylized typography or a Stable Diffusion model; (3) prompt GPT-4o to generate chain-of-thought rationales that explain the safety judgment for each instruction–image pair, together with a final answer; (4) incorporate an equal amount of purely benign data, carefully matched in format and quantity. We then propose a variant of DPO [13] to train MLLMs on this preference dataset using CoT-based safety reasoning instances to teach them to engage in safety reasoning and perform intent-aware generation.

We evaluate ReaSafe against various defense baselines on multiple datasets, using two MLLMs and four jailbreak attacks. Extensive experimental results show that ReaSafe achieves a favorable balance between safety and helpfulness. It significantly improves robustness against various jailbreak attacks without degrading general performance, while remaining sensitive to seemingly risky but benign queries. For example, ReaSafe reduces the overall attack success rate to 3.1% under advanced attacks across the three datasets, while maintaining a low refuse-to-answer rate of 9% on MOSSBench [14], substantially outperforming strong baselines.

The main contributions of our work are as follows:

1) We analyze the limitations of existing defenses (see Figure 1) and propose ReaSafe, a reasoning-centric safety alignment framework for MLLMs. ReaSafe includes a pipeline for constructing a safety-reasoning preference dataset and leverages reinforcement learning to improve user-intent inference.
2) We conduct comprehensive experiments to evaluate the effectiveness, utility, and sensitivity of ReaSafe. Evaluation results against strong baselines show that ReaSafe achieves robust defense, while maintaining or even improving the overall performance on general tasks.
3) We publicly release the safety-reasoning dataset compiled in ReaSafe to facilitate future research on safer MLLMs.

For reproducibility, our code is available in a Github repository at https://github.com/xndong/ReaSafe.

II. Method

To equip MLLMs with the capability for safety reasoning over user queries, we first construct a preference dataset whose responses align with human values and include chain-of-thought (CoT) rationales (§ II-A). We then employ reinforcement learning to align MLLMs with this dataset (§ II-B).

A. Constructing the Preference Dataset

Compiling Instruction–image Pairs. As shown in Figure 2, we begin by curating harmful instructions (*i.e.,* queries) from existing text-only datasets, primarily AdvBench [15]. To improve coverage, we further synthesize harmful queries by prompting GPT-4o with banned categories and keywords derived from OpenAI's safety guidelines. In parallel, we collect and adapt benign instructions that appear suspicious but are actually safe, sourced from XSTest [16]. These examples are carefully selected to resemble harmful queries in surface form while being semantically harmless.

To construct the visual inputs, we first filter the collected instructions to prevent data leakage by removing samples whose 5-gram Jaccard similarity with test-set instructions exceeds 0.8. We then either render the instructions in stylized typography or use a Stable Diffusion model to generate images corresponding to harmful keywords identified in the instructions. In total, we obtain over 800 text–image pairs.

Generating CoT Rationales and Preference-formatted Responses. After preparing the textual and visual inputs, we prompt GPT-4o to generate a full response consisting of a CoT rationale followed by a final answer. The prompt guides the model through three steps: (1) interpreting the visual input to extract visual semantics for downstream safety reasoning;

(2) linking the image with the textual input to infer user intent, *e.g.,* by explaining why a query is harmful or why it appears risky but is actually benign; and (3) producing a final response, *i.e.,* either a refusal or an appropriate answer. To improve data quality, we further prompt GPT-4o to evaluate its own responses on a 1–5 scale and discard those rated below 4.

For the preference-formatted dataset, the *chosen response* contains both the rationale and the final answer, whereas the *rejected response* contains only the final answer. Overall, our dataset comprises two complementary components for safety reasoning in MLLMs: (1) harmful instructions that encourage refusals of malicious queries; and (2) seemingly risky but benign instructions that improve the model's ability to infer the user's true intent, thereby mitigating over-defense.

B. Safety Reasoning via Preference Optimization

We propose a variant of Direct Preference Optimization (DPO) [13] to train a policy model π_θ, *i.e.,* the MLLM with safety reasoning capability, from a reference model π_{ref}, *i.e.,* the original MLLM. Following DPO, and based on the Bradley–Terry model, we express the human preference probability in terms of only the policy model π_θ and reference model π_{ref}, as shown in Equation 1, where y_w and y_l represent the chosen response (winner) and the rejected response (loser), respectively, and σ stands for `sigmoid` function.

$$p^\theta\left(y_w \succ y_l \mid x\right) = \sigma\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right) \quad (1)$$

In our preference formulation, the chosen response y_w contains a brief intent analysis, whereas the rejected response y_l is a direct answer without reasoning. Optimizing such preferences alone may encourage overly verbose response or indiscriminate reasoning across inputs. To mitigate this effect, we introduce *pairwise length-difference regularization* into the training objective. As shown in Equation 2, this regularizer limits the additional length introduced by safety reasoning by penalizing cases in which the chosen response exceeds the rejected response by more than a predefined margin δ.

$$\mathcal{R}_{\text{len}} = \lambda \cdot \max(0,\ |y_w| - |y_l| - \delta) \quad (2)$$

Overall, the training objective of the policy model in ReaSafe is formulated as a maximum likelihood loss for a parameterized policy π_θ, as shown in Equation 3:

$$\mathcal{L}_{\text{ReaSafe}}\left(\pi_\theta; \pi_{\text{ref}}\right) = \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[-\log p^\theta(y_w \succ y_l \mid x) + \mathcal{R}_{\text{len}}\right] \quad (3)$$

During ReaSafe training, to prevent the model from overfitting to refusal responses, which may lead to degraded general performance, we also balance the training data with an equal amount of purely benign queries, *i.e.,* those that are safe both in surface form and underlying intent. For this purpose, we employ the MMStar dataset.

III. Experiments

A. Experimental Setup

Implementation Details. We use Deepseek-VL-chat-7B [2] and Qwen2.5-VL-7B [1] as victim MLLMs in jailbreak evaluations, and fine-tune them with LoRA on our preference dataset using 4 NVIDIA RTX 4090D GPUs. The LoRA configuration uses `rank=128`, α=128, and a dropout rate of `0.1`. We train the models with a batch size of `64`, a learning rate of `2e-4`, a cosine decay scheduler, and `3` epochs.

The margin δ controls the allowed reasoning budget: a larger δ permits more verbose safety reasoning. We set δ to the 80th percentile of the empirical distribution of length differences $|y_w| - |y_l|$, yielding a reasoning budget of $\delta = 126$ tokens. We set the length-penalty coefficient to $\lambda = 10^{-3}$ so that the regularization remains secondary to the main objective.

Defense Baselines. We compare ReaSafe to several recent defenses. Specifically, we employ VLGuard [5] and SPA-VL [6] as training-based defense baselines. For prompt-based defense baselines, we include AdaShield [11], while ECSO [7] and CIDER [9] are used as detection-based defense baselines.

Jailbreak Attacks. We evaluate ReaSafe against defense baselines under four advanced jailbreaks, including FigStep, Query-Relevant-Image (QRI) [17], MML and Hades [18].

B. Defense Efficacy Evaluation.

Datasets and Metrics. We evaluate ReaSafe against defense baselines on three benchmark datasets: SafeBench [4], MM-SafetyBench [17], and HADES [18], each containing harmful queries designed to elicit unsafe responses. We adopt the attack success rate (ASR) as the evaluation metric, defined as the proportion of queries that yield jailbroken responses. Consistent with previous work, a response with a GPT-assigned harmfulness score of 5 is considered a successful jailbreak. A lower ASR indicates stronger defense performance.

Evaluation Results. As shown in Table I, ReaSafe consistently reduces the ASR across all attack settings, achieving performance that is comparable to or significantly better than that of competitive baselines. For example, ReaSafe reduces the average ASR from 36.5% to 3.1% on DeepSeek-VL-7B, and from 53.3% to 11.6% on Qwen-2.5-VL-7B, highlighting its strong defense efficacy against jailbreak attacks. On DeepSeek-VL-7B, ReaSafe attains near-zero ASR under multiple attack settings, including FigStep and MML-mirror, across both SafeBench and MM-SafetyBench. Moreover, ReaSafe significantly outperforms training-based and prompt-based defenses, including SPA-VL, ECSO, and CIDER in terms of overall defense effectiveness, and we observe consistent results on Qwen-2.5-VL-7B.

Interestingly, while VLGuard and Adashield show comparable performance in blocking harmful queries, they fail to respond appropriately with queries that appear harmful on the surface but are semantically benign (described below).

C. Avoiding Over-Refusal.

Datasets and Metrics. To evaluate whether ReaSafe mitigates over-refusal on semantically harmless queries that appear unsafe at the surface level, we employ MOSSBench [14] and safe-safe split of VLGuard [5] as benchmarks. We adopt the refusal answer rate (RAR) as the metric, defined as the

TABLE I: Comparison of ASR for FigStep, QRI, MML and Hades jailbreak attacks on two MLLMs, showing the performance of ReaSafe (our defense) vs. baselines across three datasets. Lower ASR means more effective blocking of harmful queries.

Models w. defenses	Safebench			MM-Safetybench			HADES	Overall
	FigStep	QRI	MML-mirror	FigStep	QRI	MML-mirror	Hades	
Victim: DS-VL-7B	58.7%	24%	27.3%	55.4%	37.4%	22.6%	24%	36.5%
w. VLGuard	**0%**	**0%**	2.7%	**0%**	**0.5%**	**0%**	5.3%	**0.9%**
w. SPA-VL	21.3%	4%	6.7%	34.4%	13.3%	14.4%	2.7%	14.2%
w. ECSO	49.3%	24%	26.7%	51.8%	25.6%	21%	18.7%	32.1%
w. CIDER	58.7%	21.3%	27.3%	55.4%	35.9%	22.6%	24%	35.9%
w. Adashield	**0%**	14%	30%	**0%**	11.3%	21%	**0%**	12.1%
w. **ReaSafe (ours)**	**0%**	11.3%	**0%**	**0%**	6%	**0%**	4%	3.1%
Victim: Qwen-2.5-VL	56.0%	37.3%	77.3%	46.7%	45.1%	70.3%	21.3%	53.3%
w. VLGuard	4%	**0%**	22%	**0%**	2.1%	66.2%	**0%**	13.8%
w. SPA-VL	10.7%	6%	60%	21.5%	22.6%	50.8%	6.7%	26.8%
w. ECSO	55.7%	28.0%	74.7%	46.2%	33.8%	64.6%	14.7%	48.4%
w. CIDER	56%	26.7%	65.3%	46.2%	30.3%	68.2%	6.7%	46%
w. Adashield	**0%**	**0%**	**4.7%**	1.2%	**0%**	**0%**	**0%**	**0.8%**
w. **ReaSafe (ours)**	**0%**	2.1%	27.7%	**0%**	**0%**	45.3%	**0%**	11.6%

proportion of queries that receive a refusal response. A lower RAR indicates that the model is better able to discern intent and avoid unnecessary refusals.

Evaluation Results. As shown in Table II, the prompt-based defense AdaShield performs poorly on both benchmarks, exhibiting a significantly higher RAR, which indicates a strong tendency to over-refuse under these evaluation settings. We also observe that VLGuard generalizes poorly across victim models, with its RAR increasing dramatically from 13% on DeepSeek-VL-7B to 67% on Qwen-2.5-VL-7B. In contrast, ReaSafe consistently yields a lower RAR than almost all baselines across both victim models, achieving RARs of 9% and 11% on MOSSBench for DeepSeek-VL-7B and Qwen2.5-VL-7B, respectively, which indicates that intent-aware safety reasoning helps avoid unnecessary refusals. Overall, ReaSafe demonstrates more stable safety performance across models than prior defenses.

TABLE II: RAR (↓) of ReaSafe compared with baselines.

Models	MossBench	VLGuard safe-safe
DeepSeek-VL-7B	6%	1%
w. VLGuard	13%	**1%**
w. SPA-VL	27%	**1%**
w. Adashield	70%	25%
w. **ReaSafe (ours)**	**9%**	**1%**
Qwen-2.5-VL-7B	7%	1%
w. VLGuard	67%	3%
w. SPA-VL	14%	**0%**
w. Adashield	86%	69%
w. **ReaSafe (ours)**	**11%**	2%

D. Preservation of General Utility.

Datasets and Metrics. A practical defense should also preserve the model's general-purpose capabilities. We evaluate ReaSafe against defense baselines on a suite of standard vision-language benchmarks, including visual question answering (VQA), optical character recognition (OCR), and visual reasoning tasks. Specifically, we adopt OCRBench [19], RealWorldQA [20], and MME-reasoning (MME-R) [21], respectively. The evaluation metrics follow the original benchmark settings, including normalized scores and accuracy.

Evaluation Results. Table III summarizes the performance of ReaSafe and the baselines on these benchmarks. Overall, ReaSafe outperforms the defense baselines and, in some cases, even significantly improves the general performance of the original models. For example, on OCRBench, ReaSafe achieves the best performance among all defense baselines on both DeepSeek-VL-7B and Qwen2.5-VL-7B, with normalized scores of 52.1 and 89.4, respectively. We observe similar trends on MME-R, along with only slight performance degradation on RealWorldQA. These results demonstrate that ReaSafe improves MLLM safety without sacrificing general-purpose vision-language performance, supporting the effectiveness of integrating safety reasoning through preference optimization.

TABLE III: Comparison of ReaSafe and baselines on general-purpose vision-language tasks.

Models	OCRBench *norm. score* ↑	RealworldQA *accuracy* ↑	MME-R *score* ↑
DeepSeek-VL-7B	42.8	53.85	295.71
w. VLGuard	43.5	**54.37**	302.5
w. SPA-VL	44.4	52.9	281.78
w. AdaShield	44.5	52.54	207.9
w. **ReaSafe (ours)**	**52.1**	51.37	**405.36**
Qwen-2.5-VL	88.9	67.71	620.35
w. VLGuard	88.6	**68.62**	615.71
w. SPA-VL	88.6	67.0	595.71
w. AdaShield	86.3	56.86	512.5
w. **ReaSafe (ours)**	**89.4**	65.49	**642.86**

E. Ablation Study and Analysis

Inference Time Cost Analysis. We conduct an inference time analysis to study the computational overhead introduced by explicit safety reasoning and to evaluate the effectiveness of the proposed objective-level regularization in improving inference efficiency. Specifically, we measure the end-to-end inference time of different defense baselines, ReaSafe and its variant ReaSafe*, which is trained without the pairwise length-difference regularization, on the first 50 samples of OCRBench using Deepseek-VL-7B. This experiment aims to answer two questions: (i) how much additional inference cost is incurred by incorporating reasoning, and (ii) whether the regularizer can reduce this overhead in practice.

The results are shown in Figure 3. We observe that baseline methods such as VLGuard and SPA-VL incur relatively low inference time, as they do not introduce reasoning during generation. In contrast, ReaSafe* exhibits higher inference cost, reflecting the additional computation required for intent-aware safety reasoning. Importantly, when the proposed *pairwise length-difference regularization* is applied, ReaSafe achieves a decent reduction in inference time. This demonstrates that our pairwise length-difference regularization effectively prevents excessive reasoning, improving the practicality of reasoning-based defenses.

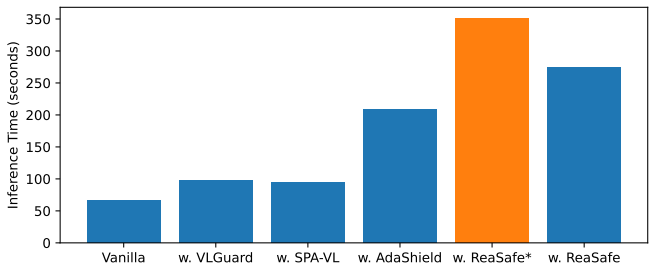

Fig. 3: Inference time comparison of different defense methods on the first 50 samples of OCRBench using DeepSeek-VL-7B. ReaSafe* denotes the variant trained without the pairwise length-difference regularization.

IV. Related Work

A. Jailbreak defenses for MLLMs.

In response to the safety vulnerabilities of MLLMs, multiple jailbreak defense techniques have been proposed.

Training-based defenses. A line of work improves safety alignment through post-training on curated datasets. VLGuard [5] constructs a vision–language safety instruction-following dataset covering various harmful categories with 2,000 samples, and fine-tunes MLLMs to learn to refuse harmful instructions. Similarly, SPA-VL [6] proposes a safety preference alignment dataset for post-training.

Our method also belongs to the post-training paradigm; however, the datasets used in VLGuard and SPA-VL are limited to refusal-style responses to harmful queries. In contrast, our dataset construction incorporates chain-of-thought safety reasoning rationales, enabling the model to learn safety reasoning from high-quality supervision. This allows the model to leverage its own reasoning capability for fine-grained intent recognition, refusing truly harmful queries while avoiding over-refusal of seemingly risky but benign ones.

Detection-based defenses. Detection-based defenses train classifiers to identify jailbreak attempts. Such methods can be feature-aware, detecting attacks based on embeddings or intermediate representations of inputs and/or outputs, or attention-aware, exploiting patterns in attention heads to recognize malicious behaviors. ECSO [7] first evaluates the safety of an MLLM's own response in a post-hoc manner and then steers generation if the output is classified as harmful. CIDER [9], SAHs [10], and ETA [8] follow a similar detection-oriented paradigm.

Prompt-based defenses. Prompt-based defenses rely on carefully designed safety prompts to guide model behavior. AdaShield [11] and Self-reminder [22] leverage manually crafted or automatically synthesized prompts to remind MLLMs to perform internal safety checks before responding.

Lastly, several works propose non-end-to-end defenses, such as BLUESuffix [23], which purifies inputs and generates a safety suffix tailored to the user query to mitigate jailbreak attempts, and SafePTR [24].

B. Jailbreak attacks on MLLMs.

Optimization-based attacks. A line of work treats MLLMs as computational systems and circumvents their safety alignment by directly optimizing malicious inputs, *e.g.,* via adversarial image perturbations or adversarial prompt suffixes [25]–[28].

Structure-based attacks. Another line of studies develops structure-based attacks that bypass MLLMs' safety checks by converting harmful textual queries into visual content, *e.g.,* via typography manipulation or text-to-image models, and then triggering malicious behavior through the visual modality. FigStep [4] converts prohibited text queries into typographic images, effectively transferring harmful intent from the textual modality to the visual modality and thereby bypassing text-only safety alignment. Query-Relevant-Image (QRI) [17] compromises MLLMs by pairing a malicious text query with a query-relevant image. MML similarly transforms queries into typographic images and further reduces surface-level toxicity by encrypting the image content; it then provides decryption hints in prompt to elicit harmful responses.

Hybrid attacks. Hybrid attacks combine the above paradigms, *e.g.,* , using structured modality transfer while additionally optimizing inputs with adversarial perturbations. HADES [18] shifts harmful intent from text to image via a text-to-image pointer and then amplifies image toxicity by appending adversarial noise. Related hybrid attack frameworks have also been proposed, including HIMRD [29].

Overall, recent jailbreak attacks increasingly emphasize structure-based and hybrid designs, which are often more effective in practice than purely optimization-based approaches.

V. Discussion

Our goal is to equip MLLMs with *safety reasoning* capability, and we therefore emphasize data quality over blindly scaling data quantity, since scaling alone often yields diminishing returns (described in LIMA [30]). We acknowledge that larger and more diverse high-quality preference datasets can improve generalization. In addition to the dataset used in the main experiments, we have compiled a larger dataset with over 3,000 instances, which significantly exceeds the scale of the 1,600-instance dataset evaluated in this work. In the current manuscript, we report results on the dataset of moderate scale that achieves a reasonable balance between safety and helpfulness, providing a reference point for the level of effort required to reach competitive performance.

We also acknowledge a potential limitation that our preference data and rationales may inherit biases or stylistic patterns from a single model. To mitigate this risk, our larger dataset is constructed by sourcing responses from multiple strong models (three in total) and aggregating their outputs during data construction, which increases stylistic and reasoning diversity and reduces over-reliance on any single model's idiosyncrasies. We leave a more systematic study on teacher diversity and dataset scaling for future work.

VI. Conclusion

This work investigates safeguarding MLLMs against jailbreak attacks. We identify that prior defenses suffer from under-safety when facing subtle jailbreak attacks and from over-refusal to harmless queries in the presence of certain visual stimuli, due to their failure to account for benign contextual intent. To address these challenges, we introduce **ReaSafe**, a novel reasoning-centric safety alignment framework. Within ReaSafe, we propose a safety-reasoning preference dataset construction pipeline. In contrast to prior works that are limited to refusal-style responses for harmful queries, our dataset incorporates chain-of-thought safety reasoning rationales. Moreover, we propose a DPO-variant algorithm that introduces a *pairwise length-difference regularization* term into the training objective, which constrains the additional verbosity induced by safety reasoning during preference optimization and prevents verbosity-driven preference exploitation. Extensive experiments demonstrate that ReaSafe consistently outperforms recent baselines by providing robust defenses against jailbreak attacks while avoiding excessive refusal and maintaining general-purpose capabilities. Overall, we conclude that explicitly harnessing a model's reasoning capability is an effective approach to addressing the complex and rapidly evolving safety threats faced by MLLMs.

References

[1] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, and K. e. a. Dang, "Qwen2.5-VL Technical Report," 2025. [Online]. Available: http://arxiv.org/abs/2502.13923

[2] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, and B. e. a. Liu, "DeepSeek-VL: Towards Real-World Vision-Language Understanding," 2024. [Online]. Available: http://arxiv.org/abs/2403.05525

[3] L. Ouyang, J. Wu, X. Jiang, D. Almeida, and C. L. e. a. Wainwright, "Training language models to follow instructions with human feedback," in *Neural Information Processing Systems*, 2022. [Online]. Available: http://arxiv.org/abs/2203.02155

[4] Y. Gong, D. Ran, J. Liu, C. Wang, T. Cong, A. Wang, S. Duan, and X. Wang. (AAAI 2025) FigStep: Jailbreaking Large Vision-Language Models via Typographic Visual Prompts. [Online]. Available: http://arxiv.org/abs/2311.05608

[5] Y. Zong, O. Bohdal, T. Yu, Y. Yang, and T. Hospedales. (ICML 2024) Safety Fine-Tuning at (Almost) No Cost: A Baseline for Vision Large Language Models. [Online]. Available: http://arxiv.org/abs/2402.02207

[6] Y. Zhang, L. Chen, G. Zheng, Y. Gao, R. Zheng, J. Fu, Z. Yin, S. Jin, Y. Qiao, X. Huang, F. Zhao, T. Gui, and J. Shao, "Spa-vl: A comprehensive safety preference alignment dataset for vision language models," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 19 867–19 878.

[7] Y. Gou, K. Chen, Z. Liu, L. Hong, H. Xu, Z. Li, D.-Y. Yeung, J. T. Kwok, and Y. Zhang. (ECCV 2024) Eyes Closed, Safety On: Protecting Multimodal LLMs via Image-to-Text Transformation. [Online]. Available: http://arxiv.org/abs/2403.09572

[8] Y. Ding, B. Li, and R. Zhang, "ETA: Evaluating Then Aligning Safety of Vision Language Models at Inference Time," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=QoDDNkx4fP

[9] Y. Xu, X. Qi, Z. Qin, and W. Wang. (2024 ACL Findings) Cross-modality Information Check for Detecting Jailbreaking in Multimodal Large Language Models. [Online]. Available: http://arxiv.org/abs/2407.21659

[10] Z. Zheng, J. Zhao, L. Yang, L. He, and F. Li. Spot Risks Before Speaking! Unraveling Safety Attention Heads in Large Vision-Language Models. [Online]. Available: http://arxiv.org/abs/2501.02029

[11] Y. Wang, X. Liu, Y. Li, M. Chen, and C. Xiao, "AdaShield : Safeguarding Multimodal Large Language Models from Structure-Based Attack via Adaptive Shield Prompting," in *Computer Vision – ECCV 2024*, vol. 15078, 2025, pp. 77–94. [Online]. Available: https://link.springer.com/10.1007/978-3-031-72661-3_5

[12] C. Zheng, F. Yin, H. Zhou, F. Meng, J. Zhou, K.-W. Chang, M. Huang, and N. Peng, "On prompt-driven safeguarding for large language models," in *Proceedings of the 41st International Conference on Machine Learning*, ser. ICML'24, 2024.

[13] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, "Direct preference optimization: your language model is secretly a reward model," in *Neural Information Processing Systems*, ser. NIPS '23, Red Hook, NY, USA, 2023.

[14] X. Li, H. Zhou, R. Wang, T. Zhou, M. Cheng, and C.-J. Hsieh, "Is your multimodal language model oversensitive to safe queries?" in *ICLR*, 2025. [Online]. Available: https://openreview.net/forum?id=QsA3YzNUxA

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

[16] P. Röttger, H. Kirk, B. Vidgen, G. Attanasio, F. Bianchi, and D. Hovy, "XSTest: A test suite for identifying exaggerated safety behaviours in large language models." Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024.

[17] X. Liu, Y. Zhu, J. Gu, Y. Lan, C. Yang, and Y. Qiao. (ECCV 2024) MM-SafetyBench: A Benchmark for Safety Evaluation of Multimodal Large Language Models. [Online]. Available: http://arxiv.org/abs/2311.17600

[18] Y. Li, H. Guo, K. Zhou, W. X. Zhao, and J.-R. Wen. (ECCV 2024) Images are Achilles' Heel of Alignment: Exploiting Visual Vulnerabilities for Jailbreaking Multimodal Large Language Models. [Online]. Available: http://arxiv.org/abs/2403.09792

[19] Y. Liu, Z. Li, M. Huang, B. Yang, W. Yu, C. Li, X.-C. Yin, C.-L. Liu, L. Jin, and X. Bai, "Ocrbench: on the hidden mystery of ocr in large multimodal models," *Science China Information Sciences*, vol. 67, no. 12, Dec. 2024.

[20] Grok-1.5 Team, "Grok-1.5 vision preview," 2024. [Online]. Available: https://x.ai/blog/grok-1.5v

[21] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, J. Yang, X. Zheng, K. Li, X. Sun, Y. Wu, R. Ji, C. Shan, and R. He, "MME: A comprehensive evaluation benchmark for multimodal large language models," in *The Thirty-ninth Annual Conference on Neural Information*

Processing Systems Datasets and Benchmarks Track, 2025. [Online]. Available: https://openreview.net/forum?id=DgH9YCsqWm

[22] Y. Xie, J. Yi, J. Shao, J. Curl, L. Lyu, Q. Chen, X. Xie, and F. Wu, "Defending ChatGPT against jailbreak attack via self-reminders," *Nature Machine Intelligence*, vol. 5, no. 12, pp. 1486–1496, Dec. 2023.

[23] Y. Zhao, X. Zheng, L. Luo, Y. Li, X. Ma, and Y.-G. Jiang, "Bluesuffix: Reinforced blue teaming for vision-language models against jailbreak attacks," in *ICLR*, 2025.

[24] B. Chen, X. Lyu, L. Gao, J. Song, and H. T. Shen. SafePTR: Token-Level Jailbreak Defense in Multimodal LLMs via Prune-then-Restore Mechanism. [Online]. Available: http://arxiv.org/abs/2507.01513

[25] L. Bailey, E. Ong, S. Russell, and S. Emmons. (ICML 2024) Image Hijacks: Adversarial Images can Control Generative Models at Runtime. [Online]. Available: http://arxiv.org/abs/2309.00236

[26] X. Qi, K. Huang, A. Panda, P. Henderson, M. Wang, and P. Mittal, "Visual Adversarial Examples Jailbreak Aligned Large Language Models," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 19, 2024. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/30150

[27] E. Shayegani, Y. Dong, and N. Abu-Ghazaleh, "Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=plmBsXHxgR

[28] Z. Niu, H. Ren, X. Gao, G. Hua, and R. Jin. (2024) Jailbreaking Attack against Multimodal Large Language Model. [Online]. Available: http://arxiv.org/abs/2402.02309

[29] M. Teng, J. Xiaojun, D. Ranjie, L. Xinfeng, H. Yihao, C. Zhixuan, and R. Wenqi. (ICCV 2025) Heuristic-Induced Multimodal Risk Distribution Jailbreak Attack for Multimodal Large Language Models. [Online]. Available: http://arxiv.org/abs/2412.05934

[30] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, S. Zhang, G. Ghosh, M. Lewis, L. Zettlemoyer, and O. Levy, "Lima: less is more for alignment," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.