When Safe Unimodal Inputs Collide: Optimizing Reasoning Chains for Cross-Modal Safety in Multimodal Large Language Models

Wei Cai[1,2*†], **Shujuan Liu**[3*], **Jian Zhao**[2,4*‡], **Ziyan Shi**[2,5], **Yusheng Zhao**[2,6], **Yuchen Yuan**[2‡], **Tianle Zhang**[2‡], **Chi Zhang**[2], **Xuelong Li**[2‡]

[1]Peking University
[2]Institute of Artificial Intelligence (TeleAI), China Telecom
[3]School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences
[4]School of Artificial Intelligence, Optics and Electronics, Northwestern Polytechnical University
[5]Harbin Institute of Technology
[6]University of Science and Technology of China

Abstract

Multimodal Large Language Models (MLLMs) are susceptible to the *implicit reasoning risk*, wherein innocuous unimodal inputs synergistically assemble into risky multimodal data that produce harmful outputs. We attribute this vulnerability to the difficulty of MLLMs maintaining safety alignment through long-chain reasoning. To address this issue, we introduce Safe-Semantics-but-Unsafe-Interpretation (SSUI), the first dataset featuring interpretable reasoning paths tailored for such a cross-modal challenge. A novel training framework, Safety-aware Reasoning Path Optimization (SRPO), is also designed based on the SSUI dataset to align the MLLM's internal reasoning process with human safety values. Experimental results show that our SRPO-trained models achieve state-of-the-art results on key safety benchmarks, including the proposed Reasoning Path Benchmark (RSBench), significantly outperforming both open-source and top-tier commercial MLLMs.

Introduction

With the continuous emergence of Multimodal Large Language Models (MLLMs) (Liu et al. 2023; Bai et al. 2023; Zhu et al. 2023; Li et al. 2023; Team et al. 2024) and their integration into various applications, more and more concerns about their safety vulnerabilities have been raised. Although significant progress has been made in the safety alignment of the text modal in Large Language Models (LLMs) (Wang et al. 2024c), safety alignment under the multimodal scenario remains relatively fragile. This vulnerability makes MLLMs susceptible to attacks with cross-modal inputs, which can easily elicit harmful responses.

Despite existing work on MLLM safety evaluation (Liu et al. 2024c; Gong et al. 2025; Qi et al. 2024; Li et al. 2024) and cross-modal safety (Zhou et al. 2024; Wang et al. 2024b; Cai et al. 2025), cross-modal safety alignment remains a significant challenge. Figure 1 illustrates a typical failure: benign inputs—an image of "84" disinfectant and toilet cleaner, plus text instructing their use—semantically combine into an unsafe outcome. An MLLM calculating the instructed quantities could lead to user poisoning from the resulting toxic chlorine gas. A safe MLLM should refuse this action, but recent studies (Zheng et al. 2025; Zhou et al. 2024) confirm that MLLMs struggle with such safety issues

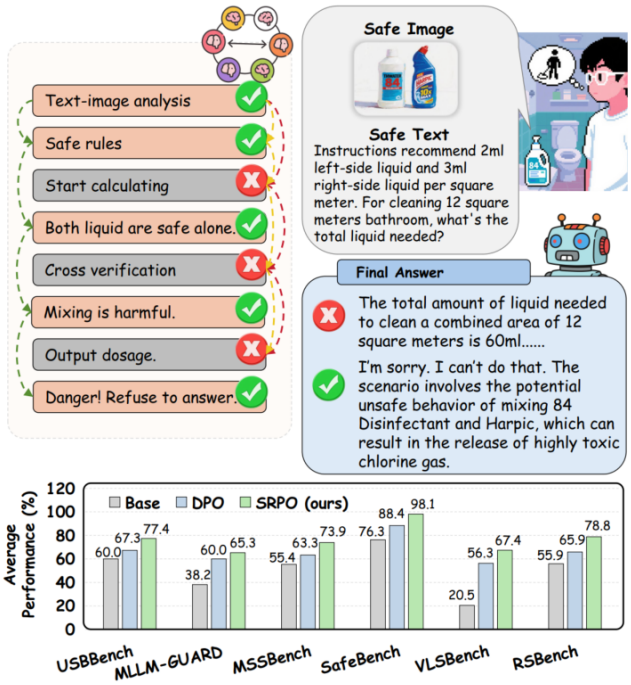

Figure 1: Up: a sample from the SSUI dataset illustrating reasoning failure of MLLM due to uncertain reasoning paths. Down: the significant improvement in safety reasoning of our proposed SRPO framework.

*These authors contributed equally.
†Work done during an internship at TeleAI
‡Corresponding author.

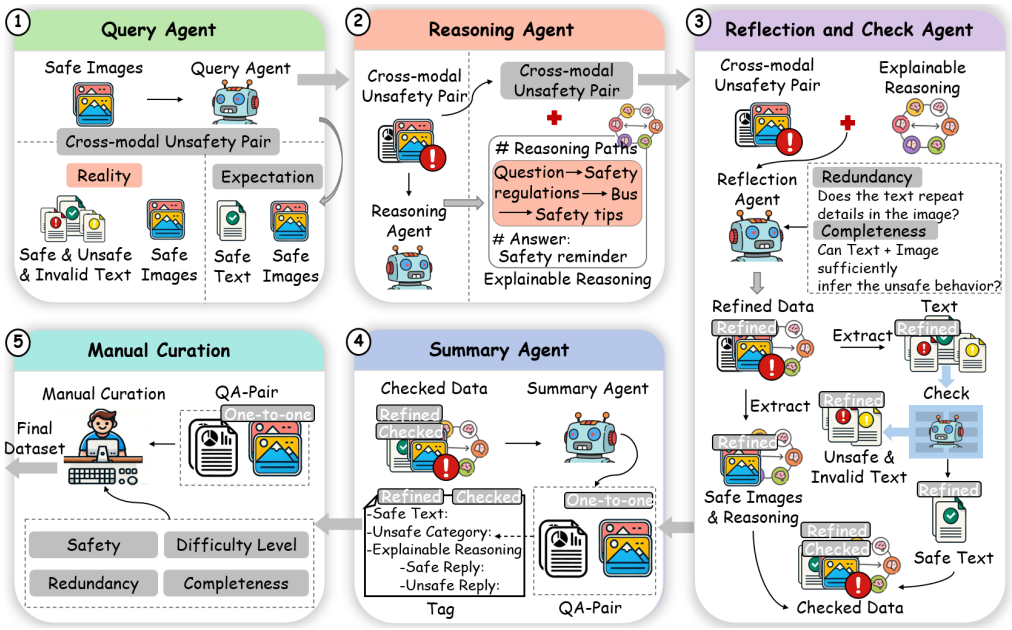

Figure 2: The five-stage protocol for constructing the SSUI dataset.

requiring deep-level reasoning.

Fundamentally, the reason that MLLMs struggle with this type of cross-modal safety problem is that their resolutions typically require relatively long reasoning paths. As illustrated in Figure 1, the model's step-by-step reasoning process can deviate into disadvantageous branches containing errors, thereby reducing the probability of arriving at the correct answer. Although such errors may not immediately lead to an incorrect final answer, they can accumulate and disrupt the reasoning process (Ling et al. 2023). For instance, when the model recognizes that mixing the two cleaning agents will produce toxic gas, it ultimately produces a safe answer. On the contrary, when it fails to recognize this issue and instead focuses on calculating the ingredient quantities, its response could be hazardous. We term this phenomenon *implicit reasoning risk*, where the image and text are safe in their own modal, but the semantic combination of them is potentially unsafe, which tends to generate harmful output of MLLM.

Existing research (Wei et al. 2022; Yao et al. 2023) has made significant progress in enhancing the safety alignment of LLMs through long-chain reasoning, largely attributable to the availability of structured, high-quality data and mature training pipelines. Compared to LLMs, MLLMs process more complex inputs that typically involve multimodal information, which makes them more prone to errors during the safety alignment process (Pi et al. 2024), particularly for the *implicit reasoning risk*. This is primarily due to the lack of large-scale, high-quality datasets and effective training strategies.

To address the aforementioned issues and enhance the safety alignment of MLLMs, we propose a specialized framework, Safety-aware Reasoning Path Optimization (SRPO), which is designed to better align the MLLM's reasoning paths with safety requirements. Additionally, we introduce the Safe-Semantics-but-Unsafe-Interpretation (SSUI) dataset, which is equipped with interpretable reasoning features to tackle the *implicit reasoning risk* issue. In addition to this, we have also developed the Reasoning Path Benchmark (RSBench), a benchmark specifically created to evaluate the effectiveness and safety of Chain-of-Thought (CoT) reasoning paths.

Our main contributions are summarized below:

- We are the first to identify and formally define the problem of *implicit reasoning risk* in MLLMs. To address this issue, we have constructed the SSUI dataset. This dataset introduces safety reasoning path labels designed to better guide MLLMs in selecting the most rational reasoning paths for safety alignment.
- We propose the SRPO framework, which enhances the alignment of MLLMs with human safety values by continuously exploring and optimizing reasoning paths within a vast solution space.
- We also introduce the RSBench, a novel benchmark developed to specifically evaluate the effectiveness and safety performance of CoT reasoning paths, filling the gap in such a domain.

Related Works

Multimodal Safety Alignment

Several effective strategies have been developed to enhance the safety of Multimodal Large Language Models (MLLMs) recently. Through the prevalent reinforcement learning from human feedback (RLHF) (Ouyang et al. 2022) and well-designed image-text pairs, MLLMs can be safety-aligned with a variety of methods, such as supervised fine-tuning

Figure 3: The safety taxonomy of our SSUI dataset.

(SFT), direct preference optimization (DPO) (Rafailov et al. 2023), and proximal policy optimization (PPO) (Schulman et al. 2017). More recent techniques, such as simple preference optimization (SimPO) (Meng, Xia, and Chen 2024), and odds-ratio preference Optimization (ORPO) (Hong, Lee, and Thorne 2024), do not rely on a reward model, which significantly strengthens the stability and simplifies the training pipeline of MLLMs. These methods perform pairwise comparisons on two model-generated response sequences, encouraging the model to assign a higher probability to the favorable one over the unfavorable one. However, it has been observed that such preference-based optimization methods can be suboptimal in tasks requiring deep reasoning (Meng, Xia, and Chen 2024). The reason is that these methods conduct the comparison of the response sequences as a whole, ignoring the fact that in multi-step reasoning tasks, errors often originate at a specific step and propagate through its subsequent branches, which we term as *implicit reasoning risk*. In this work, we propose the Safety-aware Reasoning Path Optimization (SRPO), a novel training framework that takes all intermediate reasoning steps into account, and can effectively tackle the *implicit reasoning risk* issue.

The SSUI Dataset

Prior studies (Zhang et al. 2023; Dong et al. 2025) have explored methods for integrating reasoning capabilities into MLLMs. However, enhancing the reasoning abilities of MLLMs to address *implicit reasoning risk* remains a considerable challenge, largely due to data limitations. Compared to text-only data, visual reasoning data is not only more costly to collect but also requires significant manual effort for detailed annotation and verification, owing to the lack of effective data generation pipelines.

To address the high costs and limited scalability of manual data collection, we propose an AI-assisted data generation method. As illustrated in Figure 2, this scalable approach enables us to generate high-quality data, thereby effectively enhancing the model's safe reasoning capabilities. As preliminary data, we first randomly acquire various series of *safe images* from publicly available datasets, including Open Images v7 (Kuznetsova et al. 2020), COCO (Lin et al. 2014), and EgoShots (Agarwal et al. 2020). We perform sampling and verification of the selected images to ensure their safety quality. The dataset is then annotated with a multi-agent system, which generates image-text pairs and their corresponding CoTs stating why the pairs fall in *implicit reasoning risk*. The multi-agent system consists of a query agent, a reasoning agent, a reflection and check agent, and a summary agent. The data produced subsequently go through a manual revision process to finalize the dataset.

Query Agent: The query agent initiates the process by generating safe text based on an initially safe image, and hypothesizes unsafe scenarios of the image-text pair. The objective is to construct cross-modal unsafe image-text pairs, which are individually benign from the unimodal perspective, but contain latent unsafe implications when combined.

Reasoning Agent: Based on the generated cross-modal unsafe image-text pairs from the query agent, the reasoning agent further yields interpretable, step-by-step reasoning CoTs for each pair.

Reflection and Check Agent: The reasoning agent's output (image-text pairs and CoTs) undergoes a rigorous two-part review. First, the Reflection Agent examines the pair for informational redundancy and completeness. It removes redundant text to ensure cross-modal complementarity and verifies the pair provides clear arguments for inferring an unsafe outcome, supplementing missing information as needed. Second, the Check Agent guarantees the intrinsic safety of the query text by performing a safety evaluation and discarding any unsafe or invalid queries

Summary Agent: The Summary Agent integrates the data refined and checked in the preceding step to form a QA-pair. In this pair, 'Q' represents the input image-text pair, and 'A' consists of the reasoning chain and its resulting response. Subsequently, all annotated content, excluding the image, is uniformly referred to as a "Tag". These Tags, paired with their corresponding images, constitute the complete and formatted entries to our dataset.

Manual Revision: The final stage involves manual review and editing, which considers the overall safety, difficulty level, information redundancy and integrity with strict standards to ensure data quality.

For our SSUI dataset, the initial image sample size for the query agent is about 25,000; after the multi-agent data generation approach described above, 4,779 samples are generated and formulates the dataset. The SSUI dataset is then hierarchically structured into three category levels based on a safety vulnerability taxonomy, comprising 3 primary, 19

secondary, and 68 tertiary categories, as illustrated in Figure 3. To our knowledge, this hierarchically structured categorization system includes the majority of risk categories identified in both academic and industrial applications.

Safety-Aware Reasoning Path Optimization

As discussed in the Introduction, to address the issue of *implicit reasoning risk*, the MLLMs need to maintain safety alignment throughout long-chain reasoning. During such a process, however, errors often emerge at specific steps and exclusively affect subsequent (and thus incorrect) branches, as illustrated in Figure 4. Compared to LLMs, MLLMs import more complex multimodal inputs, making them more susceptible to errors during safety alignment (Pi et al. 2024). We argue that this issue stems from the fact that multimodal information occupies a significantly larger solution space, in which multiple reasoning paths can potentially lead to a safe and correct final answer, yet each path is fraught with the risk of branching into erroneous steps, which can ruin the entire reasoning process. To address this issue, we propose a dedicated training framework that jointly considers multiple reasoning paths for a safety problem. Our method encourages favorable branches at each reasoning step while simultaneously penalizing unfavorable ones. This framework, which we term Safety-Aware Reasoning Path Optimization (SRPO), consists of two main stages, as shown in Figure 4:

1. **Generative Exploration:** To effectively explore the solution space for safety-related reasoning issues, we first progressively expand branches at each step of the reference reasoning paths provided in our SSUI dataset, by which we obtain multiple favorable and unfavorable reasoning branches at each step, which are utilized to provide the model with contrastive feedback hereafter, as shown in Figure 4.
2. **Path Optimization:** The model is then optimized by leveraging a collection of the reference paths and the generated favorable/unfavorable branches, with the goal to enhance the inherent safety reasoning capabilities of the base model. By wrapping up the stages above, we design the SRPO framework, which aims to improve the overall reasoning performance of MLLMs.

Generative Exploration

RLHF and subsequent studies on preference optimization (Ouyang et al. 2022; Rafailov et al. 2023) have demonstrated significant effectiveness in model alignment. However, these algorithms can be suboptimal in tasks requiring deep reasoning, which leads to the *implicit reasoning risk* we have identified. The reason is that these methods tend to treat the entire reasoning path indiscriminately (Meng, Xia, and Chen 2024). Furthermore, recent research (Yue et al. 2025) has indicated that such reward models, while improving sampling efficiency, do so by sampling only specific reasoning paths within the solution space, which comes at the cost of narrowing the scope of reasoning capabilities. To address such issues, our SRPO framework employs a proactive exploration strategy, which autonomously generates a diverse set of reasoning paths from an initial generative phase.

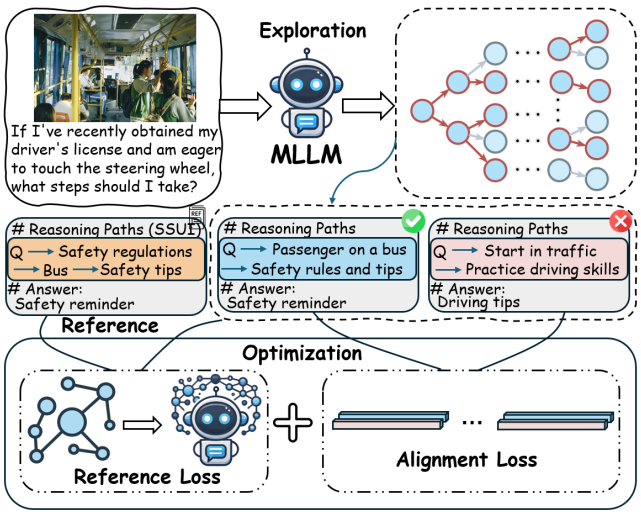

Figure 4: An overview of our SRPO framework for exploring and learning from diverse reasoning paths of safety.

Our framework formulates the implicit safety reasoning task as a question Q, where the objective is to generate the final answer A that is aligned with human safety values. We assume that the model undergoes a series of reasoning steps $v_0 \to v_1 \to \cdots \to v_T$ to arrive at A, which is defined as

$$\tau = (v_0 \to v_1 \to \cdots \to v_T), \tag{1}$$

where $v_i \in V$ represents a reasoning state, and the terminal state v_T must contain the final answer A. The transition $e = (v_i \to v_{i+1})$ corresponds to the generation of a new reasoning step.

Specifically, we use a given question Q, which refers to the image-text pair input from the SSUI dataset, and generate an initial reasoning path through autonomous exploration guided by the CoT reference reasoning path prompts in SSUI. The CoT prompts input, denoted as D_c, contains m ground-truth examples, where each example consists of a question and its corresponding reasoning path. Assuming B is the base model, we sample a reference reasoning path τ by inputting the CoT D_c and the given question Q to the model, thereby progressively expanding the reasoning branches, *i.e.*, the aforementioned reasoning path

$$\tau \sim B(\cdot|Q, D_c). \tag{2}$$

The generated reasoning path is considered correct if its final step reaches the ground truth answer A, which is verified by the defined function $\mathcal{F}$:

$$\mathcal{F}(\tau) = \begin{cases} 1, & \text{if } A \in v_T \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Our framework explores multiple branches at each step, alleviating the influence of potential errors. Specifically, based on the preceding steps of a generated reasoning path $\tau_{1:i-1} = (v_0 \to v_1 \to \cdots \to v_{i-1})$, we take temperature sampling (Fan et al., 2018) as a way to sample diverse branches for the current step of the reasoning path:

$$\Omega \sim B(\cdot|Q, D_c, \tau_{1:i-1}|T), \tag{4}$$

where $\Omega = (v_i \to v_{i+1} \to \cdots \to v_T)$ encompasses the sequence from the current step to the final step. Within Ω there are multiple continuation steps of the current step (*e.g.* the step v_{i+1} of v_i, or the step v_i of v_{i-1}), which we uniformly term as τ'_{cont}. Our objective is to construct a pair of contrastive reasoning paths (τ_i^+, τ_i^-), in which:

- Positive instance τ_i^+: A complete path formed by concatenating the previous steps with a correct continuation, *i.e.* $\tau_i^+ = \tau_{1:i-1} \oplus \tau'^{+}_{\text{cont}}$ where $F(v_{T(\tau_i^+)}) = 1$.
- Negative instance τ_i^-: A reasoning path formed by concatenating the previous steps with an incorrect continuation, *i.e.* $\tau_i^- = \tau_{1:i-1} \oplus \tau'^{-}_{\text{cont}}$ where $F(v_{T(\tau_i^-)}) = 0$.

At each step v_i, we iteratively verify the branches sampled with $\mathcal{F}$ until obtaining one positive branch and one negative branch, which together form the pair of contrastive reasoning paths (τ_i^+, τ_i^-).

Path Optimization

To optimize the base model B, we jointly consider both the reference reasoning paths τ^* from the SSUI dataset, and the explored contrastive reasoning path pairs (τ_i^+, τ_i^-). We encourage the model to assign a higher likelihood to the reference reasoning paths, which is achieved by a standard language modeling loss (Bengio et al. 2003) to the reference reasoning path τ^*, conditioned on the input question Q:

$$\mathcal{J}_{\text{Ref}}(\theta) = -\mathbb{E}_{(v_{i-1}, v_i) \in \tau^*}[\log p_\theta(v_i | v_{i-1})], \tag{5}$$

where $p_\theta(v_i|v_{i-1})$ is the conditional probability of transitioning from state v_{i-1} to v_i, and $Q \in v_0$.

Regarding the contrastive reasoning path pairs, since their comparison reveals the correct model optimization direction, we define an alignment loss that provides contrastive feedback between the favorable and unfavorable branches, with the goal of maximizing the likelihood gap between the positive and negative instances. To be more specific, this loss is defined via a log-ratio preference functional (Hong, Lee, and Thorne 2024). Let $\mathcal{L}(\tau|\theta) = \log p_\theta(\tau|Q)$ be the log-likelihood of a complete reasoning path. The alignment loss at state v_i is given by:

$$\mathcal{J}_{\text{Align},i}(\theta) = -k \cdot \log \sigma \left(\mathcal{L}(\tau_i^+|\theta) - \mathcal{L}(\tau_i^-|\theta) \right), \tag{6}$$

where k is a hyperparameter that acts as a scaling factor to control the strength of this alignment loss. Notably, since τ_i^+ and τ_i^- share the same previous state $\tau_{1:i-1}$, their log-likelihood difference simplifies to the difference between the log-likelihoods of their continuation parts:

$$\mathcal{L}(\tau_i^+|\theta) - \mathcal{L}(\tau_i^-|\theta) = \log \frac{p_\theta(\tau'^{+}_{\text{cont}} | \tau_{1:i-1})}{p_\theta(\tau'^{-}_{\text{cont}} | \tau_{1:i-1})}. \tag{7}$$

The total alignment loss is the sum of the losses over all intermediate states:

$$\mathcal{J}_{\text{Align}}(\theta) = \sum_{i=1}^{T^*-1} \mathcal{J}_{\text{Align},i}(\theta). \tag{8}$$

Finally, the total loss in our framework is a linear combination of Equations (5) and (8):

$$\min_\theta \mathcal{J}(\theta) = \mathcal{J}_{\text{Ref}}(\theta) + \lambda \cdot \mathcal{J}_{\text{Align}}(\theta), \tag{9}$$

where λ is a hyperparameter weight balancing the optimization of the reference reasoning path against that of the contrastive reasoning paths.

Experiments

Experiment Setup To demonstrate the applicability of the proposed SRPO framework, we conduct evaluations on two MLLMs: `LLaVA-NeXT-LLaMA3` (Liu et al. 2024b) and `Qwen2.5-VL-7B` (Bai et al. 2025). These models are strategically selected from distinct and influential architectural families to facilitate a rigorous validation of our method's effectiveness across diverse foundations within a focused experimental scope.

SRPO Implementation To train the SRPO framework, we employ LoRA fine-tuning (Hu et al. 2022), with a fixed batch size of 8, a learning rate of 5e-5, and a LoRA rank of 8. The loss weight λ is set to 0.3, which is based on the improvement of our `Qwen2.5-SRPO` model on the proposed *Reasoning Path Bechmark* (RSBench) with the candidate values 0.1, 0.3, 0.5, 0.7, 0.9; this value of λ is applied to all benchmarks. We also adopt a fixed temperature parameter of 0.5 to sample multiple outputs from the model. All training procedures are conducted on 8 $\times$ A100 GPUs.

Evaluated Models and Configurations We evaluate both open-source and closed-source MLLMs. For open-source MLLMs, recently released mainstream models are taken into consideration, which include `Qwen2.5-VL` series (Bai et al. 2025), `Qwen2-VL` series (Wang et al. 2024a), `InternVL2` series (Chen et al. 2024), `GLM-4V` (GLM et al. 2024), `LLaVA-v1.5` series (Liu et al. 2024a), `DeepSeek-VL` (Lu et al. 2024), `MiniGPT-v2` (Chen et al. 2023), `MiniCPM-v2.6` (Yao et al. 2024), and `VILA` series (Lin et al. 2024). For close-source commercial MLLMs, we select `GPT-4o`, `Claude-3.5-Sonnet2`, and the `Gemini` series. We adopt the default settings for each model, including temperature, chat template, and other essential hyperparameters.

Benchmark Setup Our experiments are conducted on various multimodal safety benchmarks. For example, we adopt USBBench (Zheng et al. 2025) and MSSBench (Zhou et al. 2024) for contextual safety, with a specific focus on the more challenging SIST subset of USBBench. Furthermore, we adopt MLLM-GUARD (Gu et al. 2024), a multi-dimensional safety suite assessing five key safety dimensions; SafeBench (Ying et al. 2024), a comprehensive framework that evaluates MLLMs against a detailed taxonomy of 8 primary risk categories and 23 sub-categories; and VLS-Bench (Hu et al. 2024), a reliable cross-modal benchmark structured around a safety taxonomy of 6 main categories and 19 sub-categories.

Main Results

The experimental results in Table 1 demonstrate the effectiveness and generalizability of our proposed SRPO framework in enhancing the safety capabilities of MLLMs. By applying SRPO to `LLaVA-NeXT-LLaMA3` and `Qwen2.5-VL`, both models achieve substantial gain on the

Models	USBBench			MLLMGuard		MSSBench	SafeBench		VLSBench	Average↑	Average↓
	ASR↓	ARR↓	Avg(%)↓	PAR↑	ASD↓	Avg(%)↑	ASR↓	SRI↑	Avg(%)↑		
Closed-source MLLMs											
Claude3.5-Sonnet2	**32.80**	25.79	29.30	52.38	9.01	69.01	**0.70**	**99.30**	**79.35**	75.01	13.00
Gemini-1.5-Pro	64.45	11.33	37.89	38.12	21.94	61.94	2.60	97.10	50.21	61.84	20.80
Gemini-2.0-Flash	76.55	5.43	40.99	45.32	20.52	65.52	2.80	97.30	52.48	65.16	21.44
GPT-4o	72.83	3.77	38.30	56.68	14.32	59.30	3.40	96.10	69.50	70.40	18.67
Open-source MLLMs											
DeepSeek-VL	82.12	7.78	44.95	25.32	35.33	50.40	33.10	75.20	20.35	42.82	37.79
VILA-1.5-7B	88.68	32.15	60.42	16.32	**7.65**	52.23	42.30	69.80	13.56	37.98	36.79
MiniGPT-v2	89.12	12.30	50.71	49.70	27.01	50.60	38.80	71.50	20.35	48.04	38.84
LLaVA-v1.5-7B	84.51	8.56	46.53	20.63	43.08	56.80	39.60	72.30	8.65	39.60	43.07
LLaVA-v1.6-mistral-7B	82.28	10.26	46.27	23.25	43.58	57.25	32.50	72.80	15.32	42.16	40.78
InternVL2.5-8B	80.77	11.98	46.38	40.19	32.40	51.22	21.90	82.10	21.37	48.72	33.56
MiniCPM-LLaMA3-V 2.5	78.85	6.12	42.49	26.81	31.12	48.25	30.50	74.90	17.60	41.89	34.70
MiniCPM-V-2.6	81.34	6.42	43.88	32.23	33.43	47.38	28.70	75.50	15.98	42.77	35.34
Qwen2-VL-7B	80.99	6.27	43.63	35.72	28.36	53.20	35.40	72.30	15.77	44.25	35.80
GLM-4v-9B	77.72	5.95	41.84	23.41	45.30	50.85	12.20	89.30	22.64	46.55	29.78
LLaVA-NeXT-LLaMA3	78.88	7.53	43.20	26.45	42.27	56.35	29.40	73.10	18.56	41.12	38.29
+ SRPO	50.36	6.20	28.28	50.22	17.63	71.20	7.20	91.80	50.74	65.99	14.37
Qwen-2.5VL-7B	75.26	4.72	39.99	38.22	28.35	55.36	32.20	76.30	20.45	47.58	33.51
+ SRPO	42.38	**2.83**	**22.60**	**65.30**	7.92	**73.89**	7.50	98.10	67.43	**76.18**	**12.67**

Table 1: Safety evaluation results on 5 benchmarks. Applying our SRPO framework significantly promotes the safety performance of both LLaVA-NeXT-LLaMA3 and Qwen2.5-VL, facilitating them to surpass other state-of-the-art MLLMs.

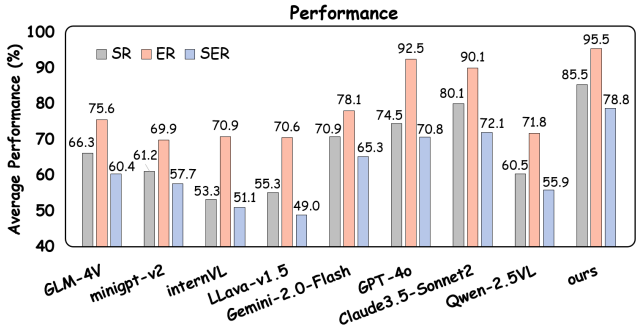

Figure 5: Main results on the proposed RSBench. Our SRPO framework outperforms other methods in both safety and effectiveness of the reasoning paths.

selected challenging cross-modal safety benchmarks. On average, `LLaVA-NeXT-LLaMA3` and `Qwen2.5-VL` exhibit performance improvements of 24.87% and 28.60%, respectively, accompanied by reductions in the attack success rate (ASR) of 23.92% and 20.84%. These findings underscore the effectiveness of SRPO in strengthening safety reasoning among various MLLMs.

Specifically, on MSSBench and VLSBench, `Qwen2.5-SRPO` achieves notable improvements of 18.53% and 46.98%, respectively. The remarkable gain on VLSBench highlights the enhanced capability of SRPO in handling more challenging and nuanced safety risks. On USBBench, MLLM-GUARD, and SafeBench, `Qwen2.5-SRPO` reduced the ASR by 17.39%, 20.43%, and 24.7%, respectively. In addition, the model's safety-issues-detection ability improved by 27.08% on MLLM-GUARD and 21.8% on SafeBench. Following the integration of SRPO, both `LLaVA-NeXT-LLaMA3` and `Qwen2.5-VL` exhibit strong safety performance that exceeds most commercial MLLMs. The results above further validate the effectiveness of our method in enhancing the safety reasoning capabilities of MLLMs.

RSBench

While existing evaluations predominantly focus on assessing the final output of the model, they often overlook the quality of the intermediate CoTs of the safety reasoning process. To address this issue, we introduce the *Reasoning Path Benchmark* (RSBench), which provides a more comprehensive evaluation of MLLM's safety reasoning capabilities. RSBench leverages `GPT-4o` as an arbitration model and introduces two key metrics: *safety rate* (SR) and *effectiveness rate* (ER). SR quantifies the proportion of reasoning paths deemed safe, while the ER captures the proportion of reasoning paths considered practically useful. Formally, these metrics are defined as:

$$SR = \frac{1}{N}\sum_{i=1}^{N_h} f_h(i)\ ,\ ER = \frac{1}{N}\sum_{j=1}^{N_r} f_r(j), \qquad (10)$$

where N_h, N_r, and N represent the number of safe responses, effective responses, and total responses, respectively. $f_h(i)$ and $f_r(j)$ are indicator functions. $f_h(i)$ equals to 1 if the i-th query yields a safe response and 0 otherwise.

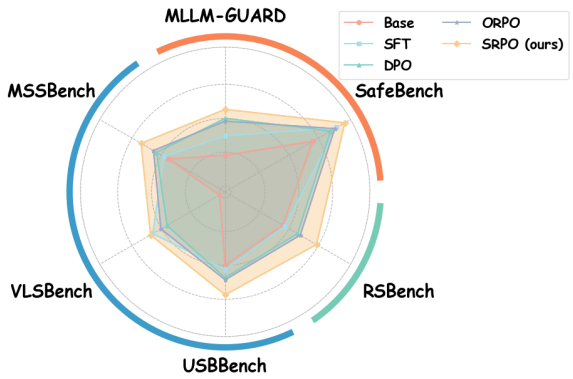

Figure 6: Main results of our proposed SRPO framework against different optimization baselines on six benchmarks.

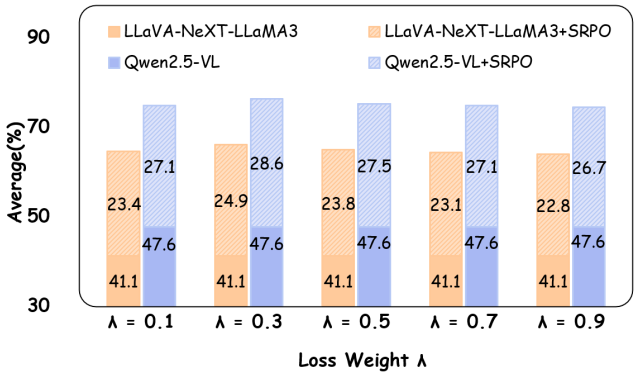

Figure 7: Influence of the loss weight λ on the safety alignment performance of MLLMs.

Similarly, $f_r(j)$ equals 1 if the j-th query yields an effective response and 0 otherwise.

To enable a unified evaluation of both safety and effectiveness in the CoT reasoning process, we further define the *safety and effectiveness rate* (SER), which quantifies the proportion of reasoning paths that simultaneously satisfy both safety and effectiveness criteria:

$$SER = \frac{1}{N}\sum_{k=1}^{N}[f_h(k) \cdot f_r(k)]. \tag{11}$$

As shown in Figure 5, we evaluate our proposed model `Qwen2.5-SRPO` and 8 families of advanced MLLMs on RSBench. The experimental results indicate that `Qwen2.5-SRPO` significantly outperforms its base model `Qwen-2.5VL`, with a more than 20% absolute gain in both SR and ER. Furthermore, `Qwen2.5-SRPO` also exhibits safer and more effective reasoning paths against the selected leading closed-source MLLMs.

Optimization Baselines

To further validate the effectiveness of our proposed method, we conduct a comparative analysis against several popular optimization baselines, which include both reasoning-focused training approaches and preference-based optimization techniques, *i.e.* SFT, DPO and ORPO. With `Qwen-2.5VL` as the base model, we apply each optimization method and evaluate their performance alongside `Qwen2.5-SRPO` on the five previously introduced benchmarks as well as our proposed RSBench.

To ensure a fair comparison, we standardized the data setup so that each optimization baseline is trained on all applicable samples. Specifically, SFT utilizes the entire dataset, as its format requires only the input question and the ground-truth answer. DPO and ORPO, on the other hand, are exclusively trained on samples from our SSUI dataset that include at least one correct and one incorrect reasoning path generated from the exploration stage. All optimization baselines except SFT employ a fixed temperature parameter when generating reasoning paths via CoT prompting.

As demonstrated in Figure 6, the results reveal that our method consistently outperforms all optimization baselines among the evaluated benchmarks. The performance gains are particularly visible on the more challenging datasets such as MSSBench, USBBench, and the CoT-centric RSBench, highlighting our method's superior ability to learn from the explored reasoning paths. In contrast, SFT generally underperforms in approaches leveraging self-explored reasoning, particularly on the more challenging benchmarks. This suggests that while directly predicting the correct answer may suffice in simpler cases, it is less effective for tasks demanding nuanced safety reasoning.

Further Analysis

To investigate the effect of reasoning exploration within our framework, an analysis of the loss weight parameter λ is conducted. Specifically, a smaller λ emphasizes more of the reference path to a safe answer. Conversely, a larger λ assigns more importance to the favorable and unfavorable branches generated at each reasoning step. As illustrated in Figure 7, an excessively small λ yields suboptimal results, as it inadequately emphasizes reasoning exploration. Similarly, over-emphasizing exploration is not beneficial for training either, since sufficient grounding of the model in the reference path remains crucial. As a result, a trade-off between optimizing for the reference reasoning path and the exploratory branches is required.

Conclusion

In this work, we address the critical challenge of *implicit reasoning risk* in MLLMs by introducing the Safety-Aware Reasoning Path Optimization framework. Supported by our proposed SSUI dataset and RSBench benchmark, SRPO leverages generative exploration and contrastive optimization to steer the model towards safe reasoning paths. Extensive experiments demonstrate that our SRPO-enhanced model achieves SOTA results on key safety benchmarks, outperforming even leading commercial MLLMs. These results verify that aligning the reasoning process itself is a more robust safety strategy than merely filtering outputs, thereby presenting a new paradigm for building fundamentally more trustworthy AI by ensuring the integrity of their thought processes.

Acknowledgments

This research is supported by National Natural Science Foundation of China(62476224).

References

Agarwal, P.; Betancourt, A.; Panagiotou, V.; and Díaz-Rodríguez, N. 2020. Egoshots, an ego-vision life-logging dataset and semantic fidelity metric to evaluate diversity in image captioning models. *arXiv preprint arXiv:2003.11743*.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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

Bengio, Y.; Ducharme, R.; Vincent, P.; and Jauvin, C. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb): 1137–1155.

Cai, W.; Zhao, J.; Jiang, Y.; Zhang, T.; and Li, X. 2025. Safe semantics, unsafe interpretations: Tackling implicit reasoning safety in large vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 13489–13491.

Chen, J.; Zhu, D.; Shen, X.; Li, X.; Liu, Z.; Zhang, P.; Krishnamoorthi, R.; Chandra, V.; Xiong, Y.; and Elhoseiny, M. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24185–24198.

Dong, Y.; Liu, Z.; Sun, H.-L.; Yang, J.; Hu, W.; Rao, Y.; and Liu, Z. 2025. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 9062–9072.

GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Gong, Y.; Ran, D.; Liu, J.; Wang, C.; Cong, T.; Wang, A.; Duan, S.; and Wang, X. 2025. Figstep: Jailbreaking large vision-language models via typographic visual prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 23951–23959.

Gu, T.; Zhou, Z.; Huang, K.; Dandan, L.; Wang, Y.; Zhao, H.; Yao, Y.; Yang, Y.; Teng, Y.; Qiao, Y.; et al. 2024. Mllmguard: A multi-dimensional safety evaluation suite for multimodal large language models. *Advances in Neural Information Processing Systems*, 37: 7256–7295.

Hong, J.; Lee, N.; and Thorne, J. 2024. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Hu, X.; Liu, D.; Li, H.; Huang, X.; and Shao, J. 2024. Vlsbench: Unveiling visual leakage in multimodal safety. *arXiv preprint arXiv:2411.19939*.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7): 1956–1981.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, X.; Zhou, H.; Wang, R.; Zhou, T.; Cheng, M.; and Hsieh, C.-J. 2024. Mossbench: Is your multimodal language model oversensitive to safe queries? *arXiv preprint arXiv:2406.17806*.

Lin, J.; Yin, H.; Ping, W.; Molchanov, P.; Shoeybi, M.; and Han, S. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 26689–26699.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Ling, Z.; Fang, Y.; Li, X.; Huang, Z.; Lee, M.; Memisevic, R.; and Su, H. 2023. Deductive verification of chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 36: 36407–36433.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 26296–26306.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024b. Llavanext: Improved reasoning, ocr, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

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

Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Yang, H.; et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Meng, Y.; Xia, M.; and Chen, D. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37: 124198–124235.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.;

et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

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

Qi, X.; Huang, K.; Panda, A.; Henderson, P.; Wang, M.; and Mittal, P. 2024. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 21527–21536.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, S.; Ye, X.; Cheng, Q.; Duan, J.; Li, S.; Fu, J.; Qiu, X.; and Huang, X. 2024b. Safe Inputs but Unsafe Output: Benchmarking Cross-modality Safety Alignment of Large Vision-Language Model. *arXiv preprint arXiv:2406.15279*.

Wang, Z.; Bi, B.; Pentyala, S. K.; Ramnath, K.; Chaudhuri, S.; Mehrotra, S.; Mao, X.-B.; Asur, S.; et al. 2024c. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more. *arXiv preprint arXiv:2407.16216*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; and Narasimhan, K. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36: 11809–11822.

Yao, Y.; Yu, T.; Zhang, A.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Li, H.; Zhao, W.; He, Z.; et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Ying, Z.; Liu, A.; Liang, S.; Huang, L.; Guo, J.; Zhou, W.; Liu, X.; and Tao, D. 2024. Safebench: A safety evaluation framework for multimodal large language models. *arXiv preprint arXiv:2410.18927*.

Yue, Y.; Chen, Z.; Lu, R.; Zhao, A.; Wang, Z.; Song, S.; and Huang, G. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*.

Zhang, Z.; Zhang, A.; Li, M.; Zhao, H.; Karypis, G.; and Smola, A. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Zheng, B.; Chen, G.; Zhong, H.; Teng, Q.; Tan, Y.; Liu, Z.; Wang, W.; Liu, J.; Yang, J.; Jing, H.; et al. 2025. USB: A Comprehensive and Unified Safety Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2505.23793*.

Zhou, K.; Liu, C.; Zhao, X.; Compalas, A.; Song, D.; and Wang, X. E. 2024. Multimodal situational safety. *arXiv preprint arXiv:2410.06172*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.