Safe in Its Own Words: Self-Guided Safety Alignment for Multimodal Reasoning Models

Adeel Yousaf[1] **Souradip Chakraborty**[2] **Mubarak Shah**[1] **Amrit Singh Bedi**[1]

[1]Institute of Artificial Intelligence, University of Central Florida [2]University of Maryland

adeel.yousaf@ucf.edu, schakra3@umd.edu,
shah@crcv.ucf.edu, amritbedi@ucf.edu

Abstract

Multimodal large reasoning (MLRMs) models can reason over images and text, but remain vulnerable to multimodal jailbreaks. Recent safety-alignment methods reduce attack success rate (ASR) by fine-tuning on safety-aware reasoning traces generated by stronger teacher models. We show that these gains come with a hidden cost of *over-refusal*, where benign inputs with sensitive visual or textual cues are incorrectly rejected. We find that external-teacher supervision induces substantial distributional shifts relative to the base model, suggesting that safety alignment depends not only on *what* behavior is learned, but also on *how* the supervision is generated. A natural alternative is to generate supervision with the base model itself. However, self-generation alone cannot correct the model's unsafe behavior; the model must be guided toward safe supervision. We find that such guidance is not straightforward. Controlling one stage of generation does not reliably control the other: safety-aware reasoning can still yield an unsafe final response, while directly controlling the final response can leave the underlying reasoning unsafe. We therefore propose **Self-Guided Safety Alignment (SGSA)**, which introduces stage-specific control over self-generated supervision. SGSA first applies a reasoning-level intervention to elicit a safety-aware trace, followed by a decision-level intervention that generates the final response conditioned on the corrected reasoning. SGSA maintains strong safety while substantially reducing over-refusal. For instance, on LLaVA-CoT, SGSA achieves ASR comparable to the corresponding external-teacher baseline while reducing average over-refusal from 71.7% to 55.9%. The project page is available here.

1 Introduction

Multimodal large reasoning models (MLRMs) achieve strong performance on complex tasks that require reasoning over both visual and textual information [Xu et al., 2025]. However, the same reasoning capability also introduces new safety risks, and MLRMs remain vulnerable to multimodal jailbreak attacks [Lou et al., 2025, Fang et al., 2025, Gong et al., 2025]. Recent safety-alignment methods address this issue by fine-tuning MLRMs on safety-aware reasoning traces generated by stronger external teacher models such as DeepSeek-R1 [Lou et al., 2025, Xia et al., 2025]. These approaches help to reduce attack success rate (ASR), but we find that this improvement comes with a hidden cost of *over-refusal*. The aligned model increasingly rejects benign inputs that contain sensitive visual or textual cues (cf. Figure 1).

This failure is particularly important for *boundary-safe* inputs, where an individual word or visual cue may appear safety-sensitive, but the complete multimodal context is benign. For example, given an image of a basketball game, the instruction "Shoot!!" should be understood as a sports-related expression rather than a violent request. Such failures are largely missed by standard utility

40th Conference on Neural Information Processing Systems (NeurIPS 2026).