

SafeLogo: Turning Your Logos into Jailbreak Shields via Micro-Regional Adversarial Training

1. Independent Defense Mechanism? (kRZ2 W1, S1; 7G1n Minor W1, S3). SafeLogo is trained with a fixed safety instructions to activate and amplify the model’s intrinsic safety alignments. The innovation of SafeLogo lies in the *visual reinforcement in the image space*. While jailbreak attacks primarily target *the textual space*, SafeLogo’s tiny-size visual defense proves highly resilient to such attacks. The effectiveness of combined defense is confirmed in Table 5 of Sect. 5, where SafeLogo with safety instructions outperforms any individual defense mechanism, including Adashield which is a *SOTA textual defense* [2].

2. Mechanistic Explanation (kRZ2 W2; RznC W1). SafeLogo is grounded in *attention mechanisms* and *cross-space defense*: 1) It is well-established that even small perturbations can significantly influence model behavior, particularly in *attention-based models*. Multiple works [3] have demonstrated how localized perturbations in both the visual and textual domains can influence the model’s output, particularly when the model relies on attention mechanisms that focus on relevant input regions. 2) Jailbreak attacks target the textual space to bypass safety mechanisms. SafeLogo’s localized visual defense is inherently difficult to circumvent, as it operates *independently from the text space*. DAVSP and ECSO have shown that defenses in a modality (visual) can complement safety measures in another modality (textual) [1, 2].

3. Benign Score Decreases (kRZ2 W3, Minor W1, S3). We acknowledge the trade-off, commonly observed in existing defenses [2, 4]. The balance can be optimized by *fine-tuning the defense weight*. As shown in Table 6 of Sec. 5, moderate defense weights (95%) achieve a more favorable balance, and is acceptable given defense performance.

4. Logo Placement (kRZ2 S2; RznC Minor W1; 7G1n W2, S2). With varied logo placements (Table 1), SafeLogo remains effective with consistently low ASR. It also shows low sensitivity to SafeLogo-harmful distance but high sensitivity to train-inference positional mismatch.

Table 1. SafeLogo Placement Sensitivity.

Setting	Train	Infer	ASR (%)	Score
No Defense	–	–	18.33	39.2
	Bottom-Right	Bottom-Right	0.67	37.7
	Bottom-Left	Bottom-Left	2.67	40.5
	Top-Right	Top-Right	3.17	37.3
	Top-Left	Top-Left	5.33	39.7
Consistent	Center	Center	3.70	36.4
	Bottom-Right	Top-Left	8.33	39.2
Inconsistent	Bottom-Right	Top-Right	1.67	38.6

5. Technical Innovation (RznC W4). Except the practical effectiveness, SafeLogo introduce several technical innovations: 1) We introduce a novel micro-regional adversarial training, which focuses on *localized defense* rather than traditional global perturbations; 2) Unlike typically exploring a single attack, SafeLogo’s *inner loop dynamically*

selects the strongest attack from a variety of jailbreak attacks, achieving stronger and more generalized defense.

6. Other Instructions (RznC W2). Experiments (Table 2) shows that SafeLogo remains robust with another fixed or dynamic safe instructions (Adashield[2]).

Table 2. Dependence on Safety Instructions.

Training	Inference	ASR (%)
No Defense	–	18.33
Fixed Instruction	Fixed Instruction	3.70
Fixed Instruction	Another Fixed Instruction	6.00
Fixed Instruction	Dynamic Instruction	2.30

7. LLM-as-Judge (RznC W3; 7G1n W1, S1). Experiment with different training and inference evaluators (Table 3) shows that SafeLogo remains consistent robust.

Table 3. Results of Different Training/Inference Evaluators.

Training Evaluator	Inference Evaluator	ASR (%)
GPT-4	GPT-4	0.33
GPT-4	LLaMA-3	2.33
LLaMA-3	GPT-4	0.33
LLaMA-3	LLaMA-3	3.00

8. Visual Perceptibility Analysis (RznC Minor W2). Quantitative evaluation using PSNR, SSIM, and LPIPS (Table 4) confirms higher visual fidelity and unobtrusiveness.

Table 4. Perceptual Quality Comparison.

Method	PSNR (dB) ↑	SSIM ↑	LPIPS ↓
SafeLogo	33.98	0.9835	0.019
DAVSP	8.17	0.1762	0.689

9. Clarification Suggestions (RznC S1, S2, S3). We fully agree that these clarifications would further strengthen the paper, and we will incorporate them in revised version.

10. Min-max Formulation (7G1n W3, S4). 1) Inner maximization $\max_{t_{\text{jail}} \in \mathcal{T}_{\text{jail}}(t_{\text{harm}})} \mathcal{L}_{\text{defense}}(\phi, x, t_{\text{jail}})$ (Eq. (4)) is approximated by selecting the strongest attack t_{jail}^* from a finite pool. The discrete selection *doesn’t require gradient backpropagation*. 2) After selecting, with the *fixed strongest attack* t_{jail}^* , gradients in Eq. (10) are back-propagated for outer loop through the differentiable components, consistent with standard AT and bilevel optimization.

References

- [1] Gou et al. Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation. In *ECCV*, 2024. 1
- [2] Wang et al. Adashield: Safeguarding multimodal large language models from structure-based attack via adaptive shield prompting. In *ECCV*, 2024. 1
- [3] Yu et al. Text-guided attention is all you need for zero-shot robustness in vision-language models. *NeurIPS*, 2024. 1
- [4] Zhang et al. Davsp: Safety alignment for large vision-language models via deep aligned visual safety prompt. *arXiv*, 2025. 1