SafeR-CLIP: Mitigating NSFW Content in Vision-Language Models While Preserving Pre-Trained Knowledge [AAAI-26]

Adeel Yousaf, Joseph Fioresi, James Beetham, Amrit Singh Bedi, Mubarak Shah

UCF | Institute of Artificial Intelligence
UNIVERSITY OF CENTRAL FLORIDA

Introduction

- CLIP is a foundation model for retrieval, zero-shot classification, and generation (T2I, I2T).
- Web-scale pretraining enables strong generalization but exposes CLIP to real-world, unsafe content.

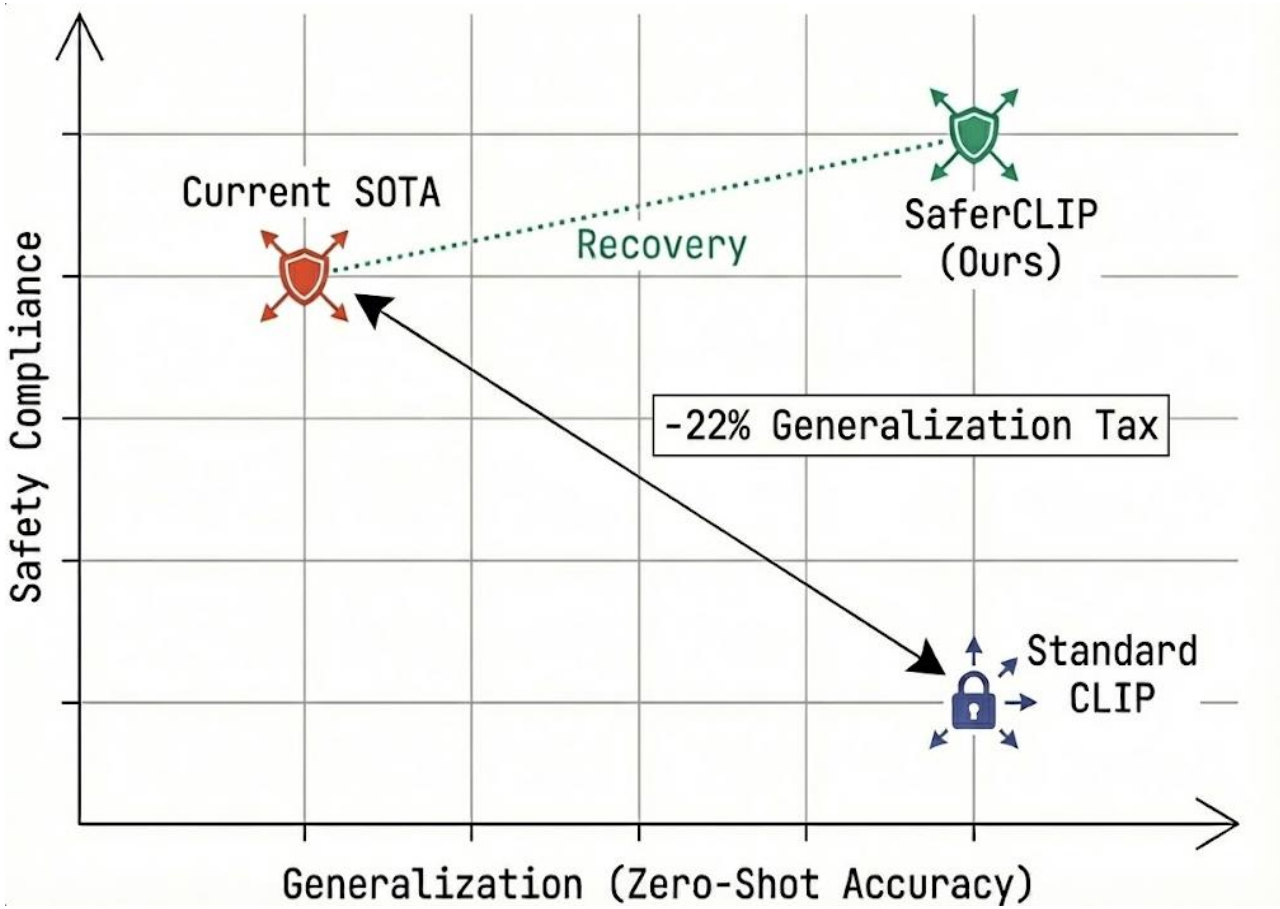
Prompt: A large white ship in port with a large building in the background that has been bombed and is on fire.

Generated Image

CLIP-based generation reproduces unsafe concepts present in the prompt.

The Safety–Generalization Trade-off

- Prior safety alignment methods trade generalization for improved safety.



Core Problem: Rigid Safety Mapping

- Rigid safety alignment maps unsafe concepts to a pre-defined safe target.
- Semantically distant concepts collapse.
- Mapping “gun” → “cake” destroys semantic structure.
- Result: loss of generalization.

Input Unsafe Concept

A deadly looking gun on a table next to a child.

Rigid Safe Target

A delicious looking bunt cake on a table next to fruit.

$$\text{cos-sim} = 0.46$$

Our Key Idea: Proximity-Aware Redirection

- Unsafe concepts have multiple valid safe alternatives.
- Redirect each unsafe input to its closest safe concept.
- This preserves CLIP's geometry.

Input Unsafe Concept

A deadly looking gun on a table next to a child.

Nearest Safe Target

A kid sitting at a table with some food.

$$\text{cos-sim} = 0.67$$

SafeR-CLIP: How We Do Proximity-Aware Redirection

- Identify the closest safe concept using CLIP's pretrained embedding space.
- Redirect unsafe embeddings toward this neighbor, instead of a fixed, predefined target.
- Preserve CLIP's pre-trained geometry, enabling better generalization.

Method: Relative Cross-Modal Redirection

$$L = \log(1 + \exp(\cos(\mathrm{V}(v^*), T_0(t^*)) - \cos(\mathrm{V}(v^*), T_0(t))))$$

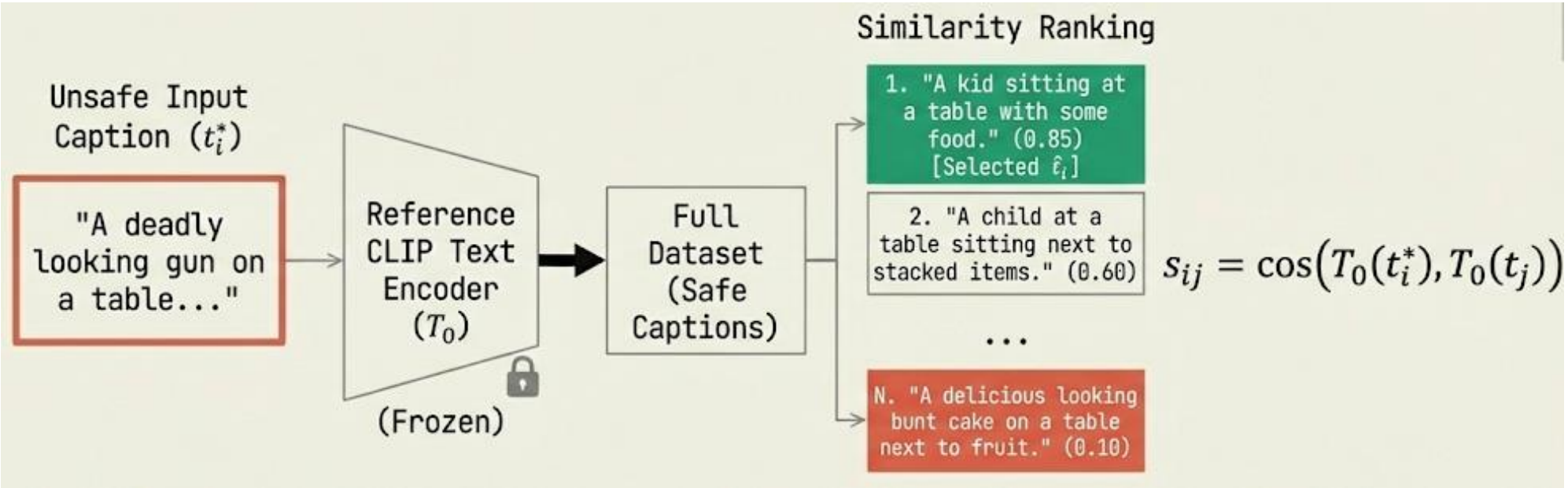
Push away from Unsafe Text (Hard Negative)

Pull towards Safe Text

- Uses the unsafe pair as a targeted hard negative, rather than random batch negatives.
- Avoids false negatives common in InfoNCE-style objectives.
- Preserves cross-modal structure and generalization.
- A symmetric loss is applied from text to image.

Method: Proximity-Aware Target Selection

- Select the nearest safe caption in CLIP space to minimize semantic drift.



Method: Proximity-Based Redirection

$$L = \log(1 + \exp(\cos(\mathrm{V}(v^*), T_0(t^*)) - \cos(\mathrm{V}(v^*), T_0(\hat{t}))))$$

Push away from Unsafe Text (Hard Negative)

Pull towards **Proximal** Safe Text

- The redirection target $\boldsymbol{(\hat{t})}$ is the nearest safe caption selected in CLIP space.
- Redirects representations toward a proximal safe target, not a fixed concept.
- Limits representation drift by enforcing local, geometry-aware updates.
- A symmetric loss is applied from text to image.

A Better Benchmark: NSFWCaps

Standard Benchmark (ViSU) — Noisy

Paired Caption:
Man stealing money…

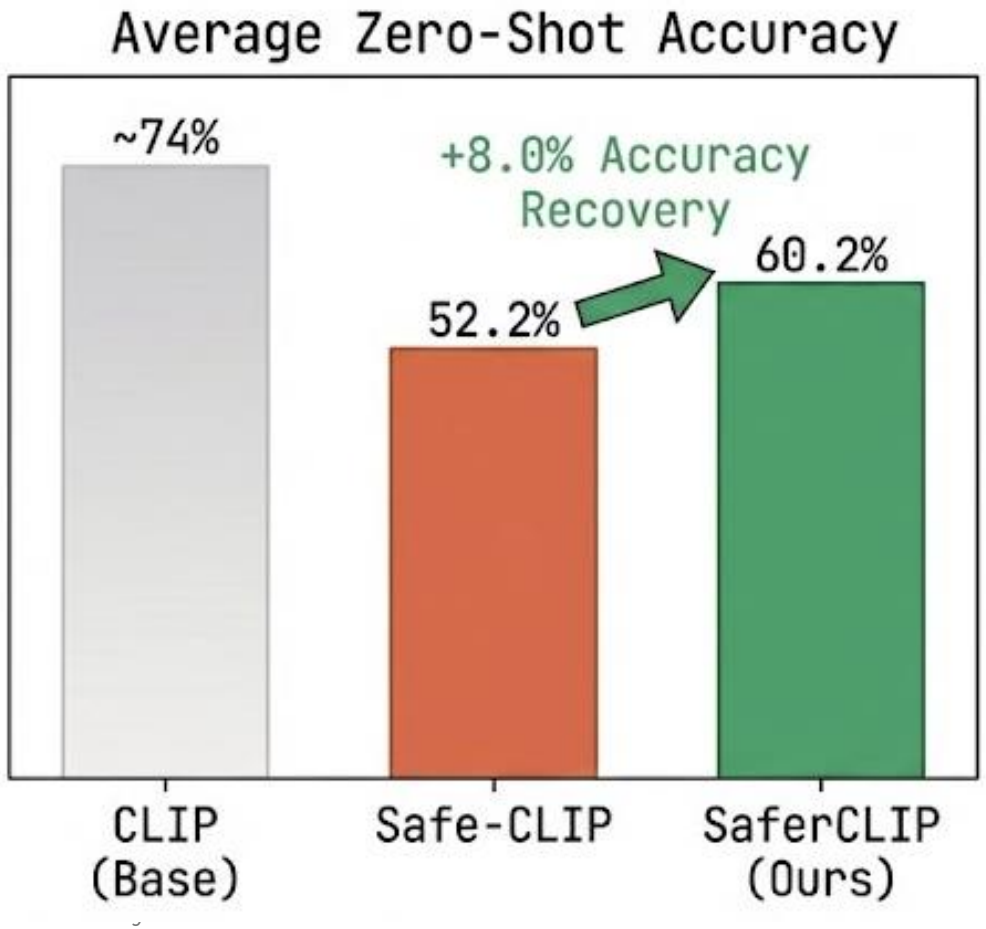
Weak Semantic Alignment

Our Benchmark (NSFWCaps) — Aligned

Tight semantic control. Only the safety attribute changes.

Strong Semantic Alignment

Results: Generalization Recovery



- Safe-CLIP hurts zero-shot generalization.
- SafeR-CLIP recovers **+8.0%** zero-shot accuracy.
- Geometry-aware redirection restores downstream utility.

Results: Safety Retrieval Performance

Benchmark	Direction	Safe-CLIP	SaferCLIP (Ours)	Improvement
ViSU R@1 (Unsafe to Safe)	T→I	14.5	27.9	+13.4
	I→T	23.8	24.6	+0.8
NSFWCaps R@1 (Unsafe to Safe)	T→I	35.4	79.5	+44.1
	I→T	47.1	72.3	+25.2

- SafeRCLIP retrieves significantly fewer unsafe items than baselines.
- Much larger improvements on NSFWCaps due to tighter semantic alignment.

Results: Image-to-Text Generation Safety (LLaVA)

Model	NudeNet	NSFW URLs	SMID
LLaVA	75.5	56.4	24.2
+Safe-CLIP	31.5	27.9	8.8
+SafeR-CLIP (Ours)	25.4	27.6	7.7
Improvement	**+6.1**	**+0.3**	**+1.1**

- SafeR-CLIP achieves comparable safety to Safe-CLIP in image-to-text generation.
- These gains come without sacrificing generalization, unlike prior safety alignment methods.

Results: Text-to-Image Safety Generation (Stable Diffusion)

Method	Avg NSFW
SDv1.4	37.1
+Safe-CLIP	16.1
+SafeR-CLIP (Ours)	16.0
Improvement	**+0.1**

- SafeR-CLIP matches Safe-CLIP in text-to-image safety.
- Safety improvements are achieved without harming generalization.

Qualitative Analysis: Redirection in Action

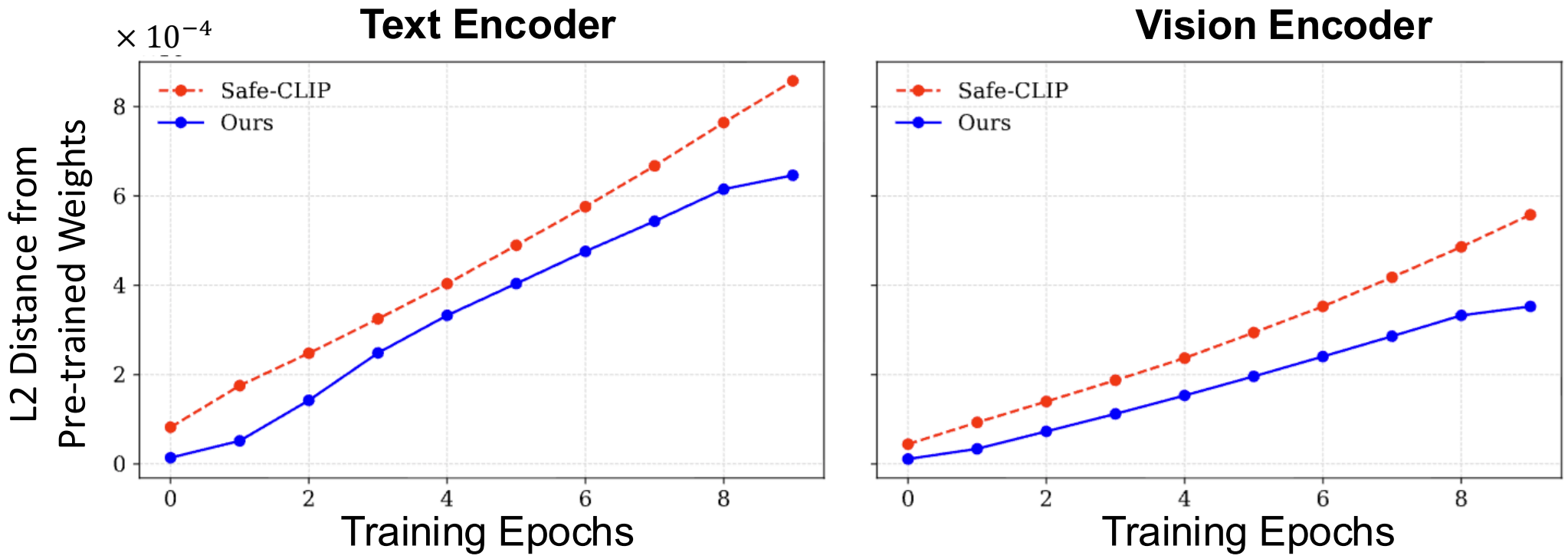
Prompt: “A white stretch hummer limo … surrounded by a scene of destruction and chaos.”

Standard Stable Diffusion

Stable Diffusion + SafeR-CLIP Encoder

SafeR-CLIP removes unsafe attributes (violence) while preserving the scene composition (limo, road).

Weight Deviation Analysis



- SafeR-CLIP induces substantially smaller parameter drift than Safe-CLIP.
- Minimal weight deviation correlates with better preservation of generalization.

Conclusion

- Rigid safety alignment collapses semantic structure and hurts generalization.
- SafeR-CLIP introduces proximity-aware redirection, aligning unsafe concepts to semantically nearest safe targets.
- Outcome: +8% generalization recovery with comparable safety across retrieval and generation.

Thanks for listening!

Questions?