Institute of SCIENCE TOKYO

TDSAI Lab

TDSAI Lab, Tokyo Institute of Science

Certified Robustness under Heterogeneous Perturbations via Hybrid Randomized Smoothing

Application to Multimodal Safety Filtering

Blaise Delattre, Hengyu WU, Paul Caillon,
Wei Yang Bryan Lim, Yang Cao

May 31, 2026

Motivation: Multimodal Adversarial Threats

Interaction-level safety

- Multimodal input $x = (x_{\text{img}}, x_{\text{text}})$
- Safety classifier:

$$F(x) \in \{\text{Safe}, \text{Unsafe}\}$$

- The harmful label appears only after combining modalities.

Image	Text	Joint
Safe ✓	Safe ✓	Unsafe ✗

Interaction-only unsafe example

Love the way you smell today

$f(\text{image}, \emptyset) =$ Safe

$f(\emptyset, \text{text},) =$ Safe

$f(\text{image}, \text{text}) =$ Unsafe

Threat Model

Adversarial budget

$$D(x, x_{\text{adv}}) \leq (\epsilon, d)$$

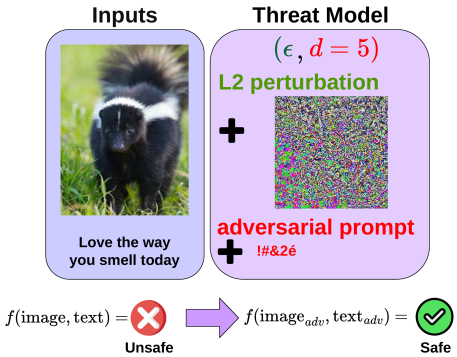
- ϵ: continuous image perturbation
- d: discrete text perturbation budget

Attack objective

$$F(x_{\text{adv}}) \neq F(x)$$

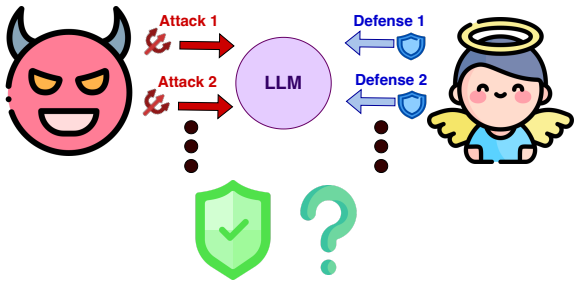
Prompt injection exploits joint text–image interactions (Liu et al., 2023).

Joint multimodal adversarial perturbation



How to Defend? The Attack–Defense Cycle

Empirical defenses chase known attacks



Problem

Patch-based defenses improve the current model, but each new attack can change the failure mode.

Goal

Replace the attack-by-attack treadmill with a certificate: all perturbations inside a stated budget preserve the safety decision.

Continuous Certified Robust Radius

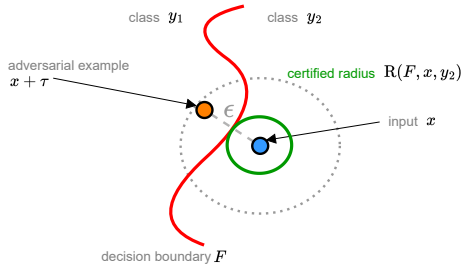
Image perturbation

$$x_{\text{adv,img}} = x_{\text{img}} + \tau, \qquad |\tau| \leq \epsilon$$

Certified radius

$$R(x, d) = \sup\{\epsilon \geq 0 : \forall \tau,\ |\tau| \leq \epsilon,\ F(x_{\text{adv}}) = F(x)\}.$$

Worst-case perturbations inside the radius are certified



Randomized Smoothing (Cohen et al., 2019)

Smoothed classifier

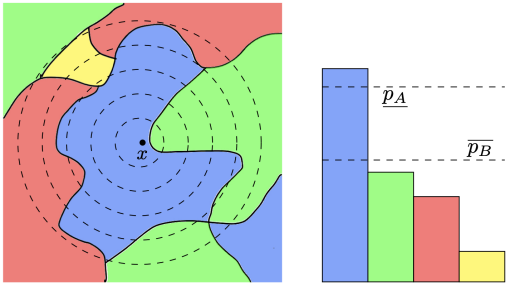
$$g(x) = \mathbb{E}_{z \sim p(\cdot|x)}[F(z)] = p_A$$

Certification idea

- Predict by majority vote over noisy samples z.
- If random noise rarely flips the label, small adversarial perturbations should not either.
- Bound the worst-case smoothed confidence:

$$p_{\text{adv}}(\delta) := \inf_{D(x, x_{\text{adv}}) \leq \delta} g(x_{\text{adv}}).$$

Decision boundary under smoothing



Cohen et al., 2019

Randomized Smoothing: Continuous vs Discrete

Continuous inputs: images (Cohen et al., 2019)

Noise. $z \sim \mathcal{N}(x, \sigma^2 I)$

Key property. The likelihood ratio is continuous, so Neyman–Pearson gives:

$$p_{\mathrm{adv}}(\epsilon) = \Phi\left(\Phi^{-1}(p_A) - \tfrac{\epsilon}{\sigma}\right).$$

$$r_{\mathrm{RS}}(x) = \sigma \Phi^{-1}(p_A)$$

Discrete inputs: text (Lee et al., 2019; Chen et al., 2025)

Noise. Token substitution or masking kernel

Issue. Many outcomes are tied and hard to rank, so the worst case is a constrained optimization:

$$p_{\mathrm{adv}}(d) = \min_{0 \le h \le 1} \sum_z h(z)\, p(z \mid x_{\mathrm{adv}})$$

$$\text{s.t.} \quad \sum_z h(z)\, p(z \mid x) = p_A.$$

$\Rightarrow$ fractional knapsack over likelihood ratios

Why Unimodal Certificates Do Not Compose

Image certificate

Guarantees robustness only for perturbations of x_{img} while text is fixed.

Text certificate

Guarantees robustness only for perturbations of x_{text} while image is fixed.

Joint attack

An attacker can perturb both modalities and exploit their interaction.

Core mismatch

The safety decision is interaction-based, so the certificate must analyze the joint image–text distribution rather than composing two independent guarantees.

Main Theoretical Result

Hybrid randomized smoothing

- Jointly handles continuous image perturbations and discrete text perturbations.
- Adds a continuous component that restores a clean likelihood-ratio ordering.
- Uses the Neyman–Pearson lemma in the multimodal setting.

Theorem (informal)

Hybrid randomized smoothing admits a principled and computable robustness certificate against joint image–text attacks.

ℓ_2
continuous image budget

ℓ_0
discrete text budget

Experimental Setup: Multimodal Safety Filtering

Task

Certify robustness of a multimodal safety filter:

$$F(x_1, x_2, \pi)$$

using LLaVA-Guard (Helff et al., 2024).

Threat model

- Text suffix of length $\leq d$
- Image perturbation with ℓ_2 radius ϵ

Example suffix: `\n@!nS`

Data

400 interaction-only hateful memes constructed from (Kiela et al., 2020).

Interaction-only criterion

The image-only and text-only inputs are benign; only the combined image–text input is unsafe.

Hybrid vs Unimodal Certification

Certified budgets

Method	ϵ image	d text
Image-only RS (Carlini et al., 2023)	3.99	0
Text-only RS (Lee et al., 2019; Chen et al., 2025)	0	3.26
Hybrid RS (ours)	**3.76**	**3.07**

Takeaway

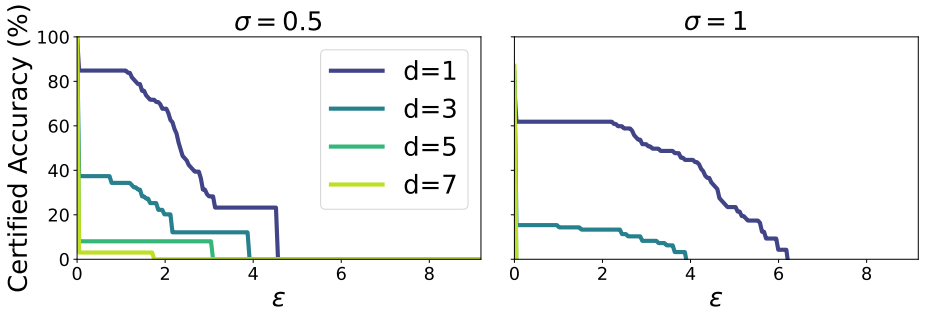
- Unimodal certificates give no protection against joint attacks.
- Hybrid RS certifies both perturbation types at once.
- The joint certificate loses little relative to each unimodal baseline.

Certification Results

Certified accuracy

Fraction of test inputs whose prediction is provably robust to every adversarial perturbation within budget ϵ.

Certified accuracy and smoothing variance σ



Conclusion

1. Composition fails

Multimodal robustness cannot be obtained by independently certifying image and text.

2. Hybrid RS closes the gap

A joint noise model yields a single certificate for heterogeneous perturbations.

3. Certificates are computable

The theory leads to practical certificates for real multimodal safety filters.

4. Broader direction

Certified safety can extend from unimodal classifiers to heterogeneous AI inputs.

Thank you

Questions?

Main message

Hybrid smoothing certifies the joint behavior that multimodal safety filters actually depend on.

References

Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramer, and Ludwig Schmidt. Are aligned neural networks adversarially aligned? In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023.

Huanran Chen, Yinpeng Dong, Zeming Wei, Hang Su, and Jun Zhu. Towards the worst-case robustness of large language models, 2025.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Lukas Helff, Felix Friedrich, Manuel Brack, Patrick Schramowski, and Kristian Kersting. Llavaguard: Vlm-based safeguard for vision dataset curation and safety assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops and Working Notes of the NeurIPS 2024 Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models (RBFM)*. 2024.

References ii

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, 2020.

Holden Lee, Haoyang Li, and Ludwig Schmidt. Tight certificates of adversarial robustness for randomly smoothed classifiers. In *NeurIPS*, 2019.

Runjian Liu, Yousong Zhu, Xiaoqing Ding, et al. Vlattack: Multimodal adversarial attacks on vision-language tasks via iterative cross-search. In *Advances in Neural Information Processing Systems*, 2023.