MITSUBISHI ELECTRIC
Changes for the Better

MITSUBISHI ELECTRIC RESEARCH LABORATORIES

ICLR

Directional Embedding Smoothing for Robust Vision Language Models

Ye Wang, Jing Liu, Toshiaki Koike-Akino
Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139, USA.

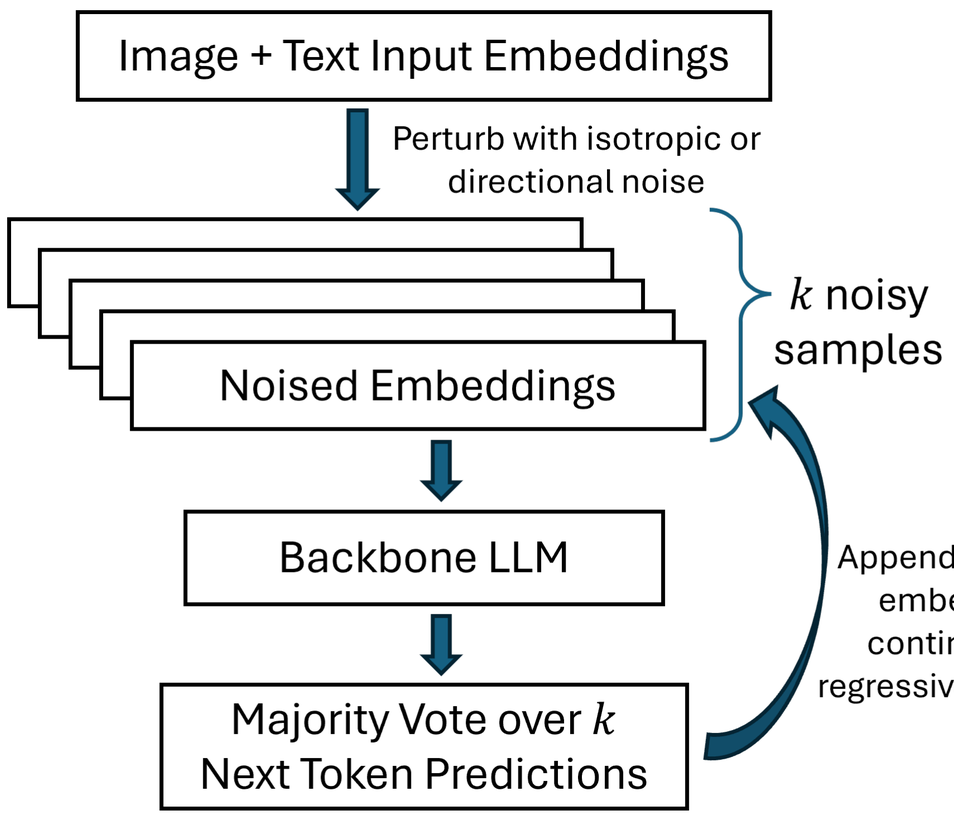
Highlights

- **Problem:** Vision-Language Models (VLMs) remain vulnerable to *multimodal jailbreak attacks*, despite safety alignment.
- **Approach:** Extend **RESTA (Randomized Embedding Smoothing & Token Aggregation)** to VLMs as a lightweight, inference-time defense.
- **Key Insight:** *Directional embedding noise* preserves semantics while disrupting jailbreak attacks.
- **Results:** Reduce attack success rate (ASR) by $\sim 2\times$ with minimal utility loss.
- **Takeaway:** Directional noise is critical for achieving strong safety-utility tradeoffs.

arXiv

RESTA: Inference-Time Defense Method

Randomized Embedding Smoothing and Token Aggregation



Investigated two types of embedding noise:

- **Normal (isotropic) noise:** iid Gaussian, zero-mean, standard deviation σ.
- **Hard (directional) noise:** aligned with each embedding vector $e \in \mathbb{R}^d$, i.e.,

$$e + \frac{ze}{\|e\|_2},$$

where z is zero-mean, scalar Gaussian noise with standard deviation σ.

RESTA Algorithm

Inputs:

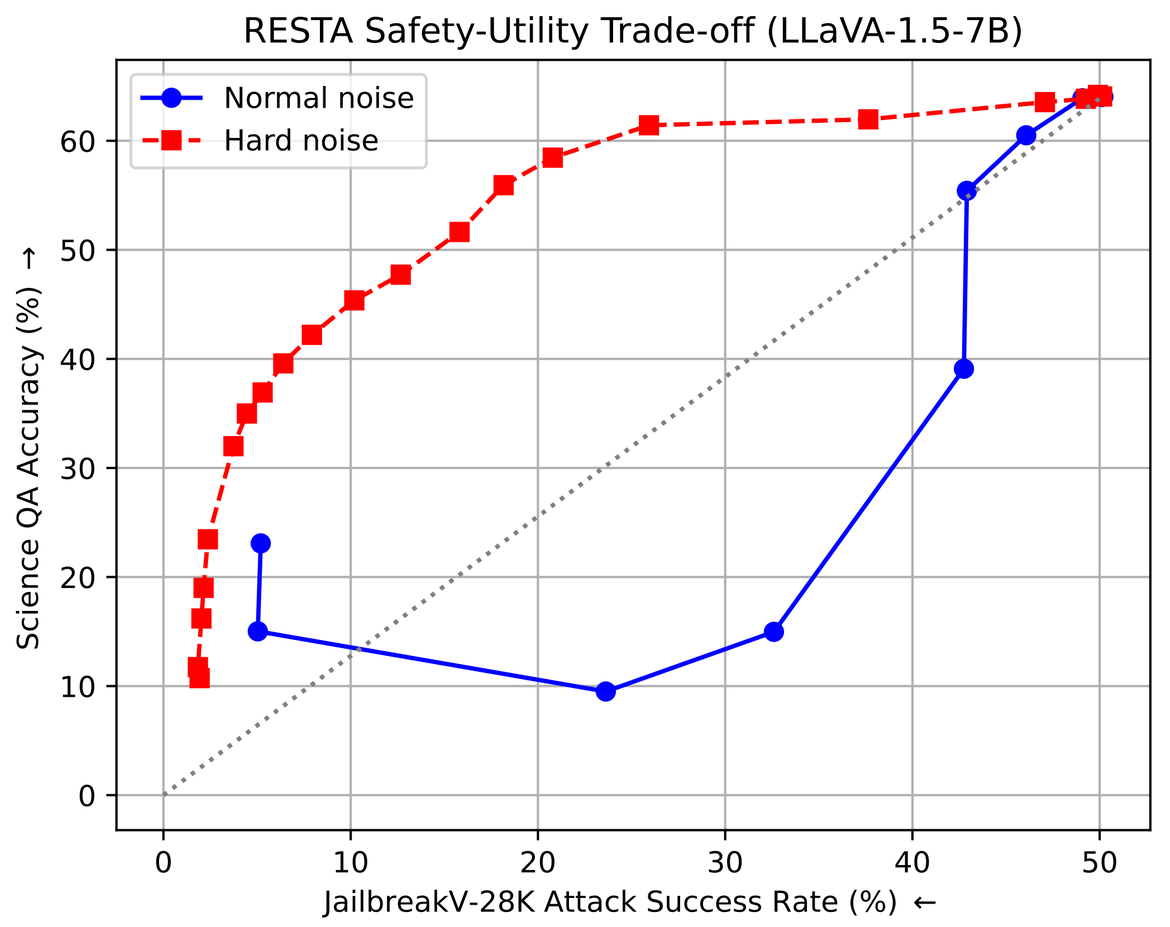
- unified embedding sequence: $\boldsymbol{e} := (e_1, \ldots, e_n) \in \mathbb{R}^{d\times *}$
- token embedding function: $E : \mathcal{X} \to \mathbb{R}^d$
- underlying LLM transformer: $f : \mathbb{R}^{d\times *} \to \mathbb{R}^{|\mathcal{X}|}$
- embedding noise (normal/hard) function: $H_\sigma : \mathbb{R}^{d\times *} \to \mathbb{R}^{d\times *}$
- number of smoothing samples: k
- maximum number of output tokens: m.

Initialize empty output sequence: $\boldsymbol{y} \leftarrow ()$.
Perturb embeddings: for $i \in \{1, \ldots, k\}$, $\tilde{\boldsymbol{e}}^i \leftarrow H_\sigma(\boldsymbol{e})$.
repeat
 for $i \in \{1, \ldots, k\}$ **do**
  Calculate next token logits: $f(\tilde{\boldsymbol{e}}^i)$.
  Select next token: $\tilde{y}^i \leftarrow \arg\max_{j\in\mathcal{X}} f(\tilde{\boldsymbol{e}}^i)[j]$.
 end for
 Majority vote: $y \leftarrow \mathrm{mode}(\tilde{y}^1, \ldots, \tilde{y}^k)$.
 Append token to output: $\boldsymbol{y} \leftarrow (\boldsymbol{y}, y)$.
 Embed token and append: for $i \in \{1, \ldots, k\}$, $\tilde{\boldsymbol{e}}^i \leftarrow (\tilde{\boldsymbol{e}}^i, E(y))$.
until $y =$ [End of Sequence token] **or** $\mathrm{length}(\boldsymbol{y}) = m$.
return Output sequence: $\boldsymbol{y}$.

Safety-Utility Tradeoff Results

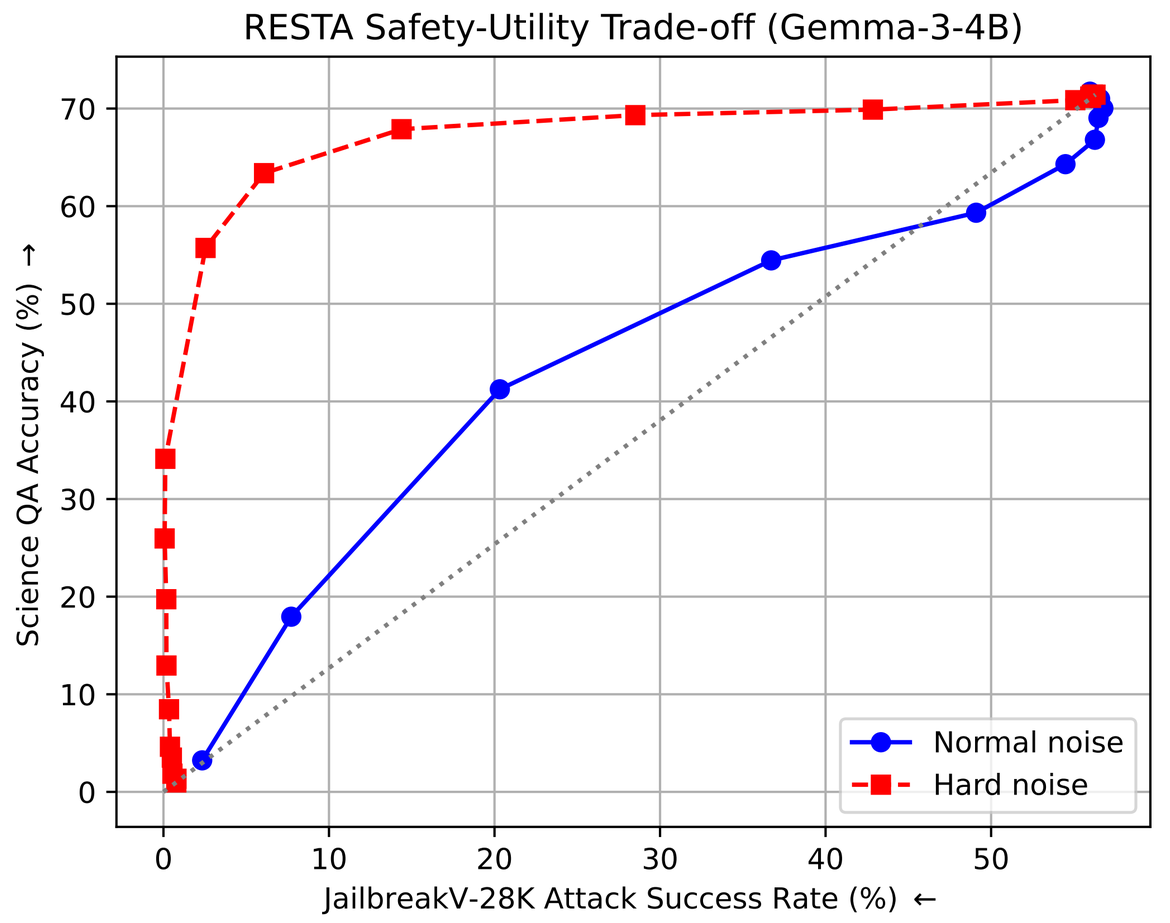
Experiments evaluated RESTA applied to LLaVA-1.5.-7B and Gemma-3-4B VLMs

- Safety evaluated with JailbreakV-28K benchmark (ASR % ↓)
  - Automatic evaluation with Llama-Guard-3-8B
- Utility evaluated with ScienceQA benchmark (accuracy % ↑)



For LLaVA-1.5.-7B, with hard directional noise ($\sigma = 0.005$):

- Jailbreak ASR reduced by $\sim 2\times$: $50.13\% \longrightarrow 25.93\%$
- ScienceQA accuracy minimally impacted: $64.07\% \longrightarrow 61.42\%$



For Gemma-3-4B, with hard directional noise ($\sigma = 0.425$):

- Jailbreak ASR reduced by $\sim 2\times$: $56.31\% \longrightarrow 28.51\%$
- ScienceQA accuracy minimally impacted: $71.42\% \longrightarrow 69.32\%$

Discussion & Takeaways

- RESTA achieves a strong **safety-utility tradeoff** for VLMs.
- **Direction matters:** aligned noise outperforms isotropic noise.
- Suggests jailbreaks exploit **fragile directions** in embedding space.
- Lightweight, **test-time defense** with no retraining required.
- Complements existing **multi-layer safety frameworks**.
- Future work: adaptive attacks, additional models/benchmarks, and investigate deeper theoretical understanding.

ICLR 2026 Workshop on Agents in the Wild: Safety, Security, and Beyond, April 26, 2026, Paper ID 190, **Email:** {yewang, jiliu, koike}@merl.com