Tübingen AI Center

EPFL

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Adversarially Robust CLIP Models Can Induce Better (Robust) Perceptual Metrics

Francesco Croce* Christian Schlarmann* Naman Deep Singh* Matthias Hein

Perceptual Similarity Metrics

- Function **sim**(x_1, x_2) that outputs a **similarity score** for a pair of images
- Encode similarity of images **as perceived by humans**

  → Capture **high-level** semantics
- Can be building blocks for various downstream systems, e.g. content filtering

Perceptual Similarity Metrics

- Early approaches: **algorithmical** (*PSNR, SSIM*)

  → unable to capture high-level semantics
- Nowadays (LPIPS [1]): With **vision encoder** ϕ, compute the similarity of images as

$$\texttt{sim}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \left\langle \frac{\phi(\boldsymbol{x}_1)}{\|\phi(\boldsymbol{x}_1)\|_2}, \frac{\phi(\boldsymbol{x}_2)}{\|\phi(\boldsymbol{x}_2)\|_2} \right\rangle$$

- ϕ could be derived e.g. from CLIP, DINO

[1] Zhang et al., The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, CVPR 2018

NIGHTS Dataset

Two Alternatives Forced Choice (2AFC) Task

1 Reference 2

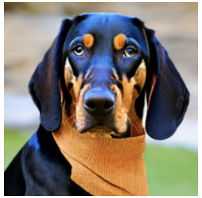
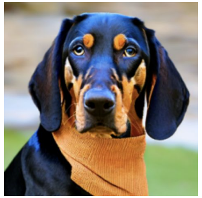
[1]

“Is 1 or 2 more similar to Reference?”

Quantifies alignment with human perception

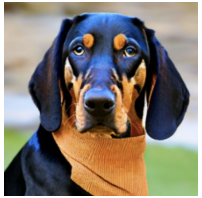
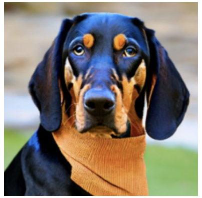
[1] Fu et al., DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data, NeurIPS 2023

Perceptual Metrics are Vulnerable

sim( , ) **>** sim( , ) [1]

[1] Ghazanfari et al., LipSim: A Provably Robust Perceptual Similarity Metric, ICLR 2024

Perceptual Metrics are Vulnerable

sim( , ) < sim( , ) [1]

+ imperceptible noise

+ imperceptible noise

⇨ **Security risk**

Goal: adversarially robust perceptual metric with high clean performance

[1] Ghazanfari et al., LipSim: A Provably Robust Perceptual Similarity Metric, ICLR 2024

Mitigation: Use robust vision encoders

- Adversarially robust vision encoders could yield robust perceptual metrics
- Our **robust fine-tuning** scheme from prior work: **FARE** [1]

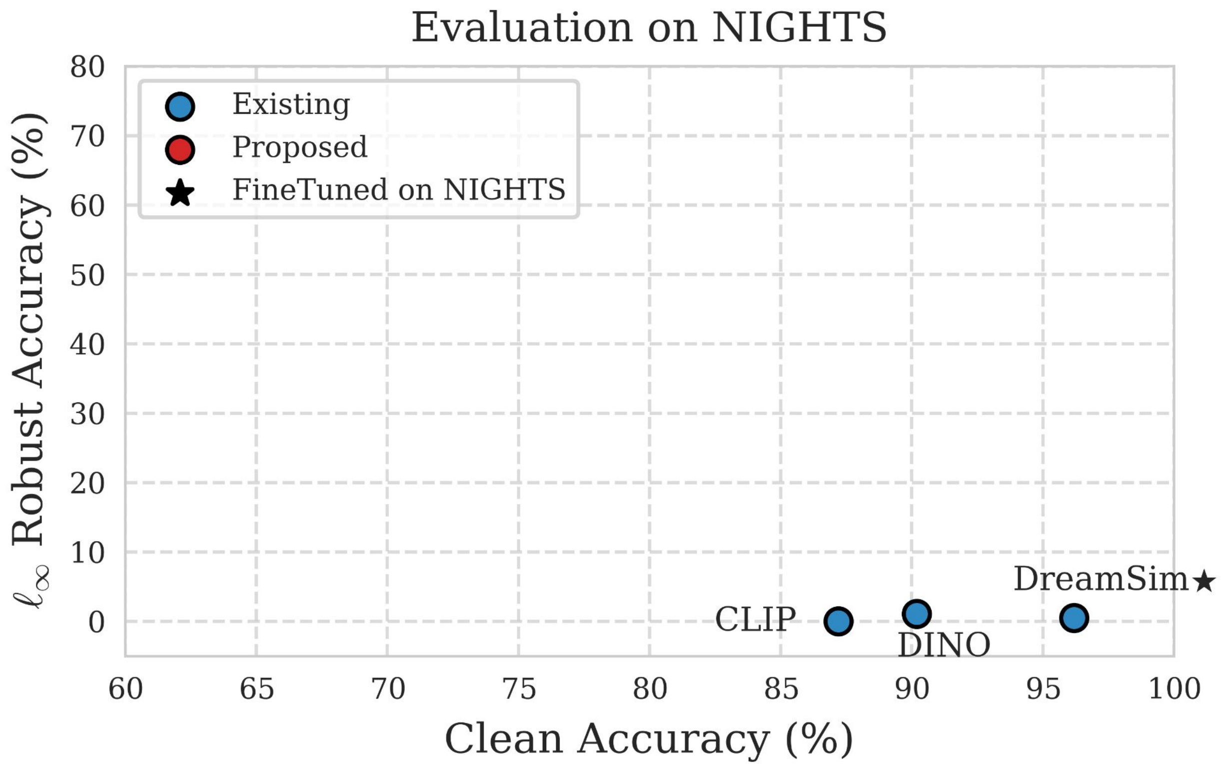
$$L_{\text{FARE}}(\phi, x) = \max_{\|z-x\|_\infty \leq \varepsilon} \|\phi(z) - \phi_{\text{Org}}(x)\|_2^2$$

$\Rightarrow$ **Ensures stability of embeddings under adversarial perturbation**

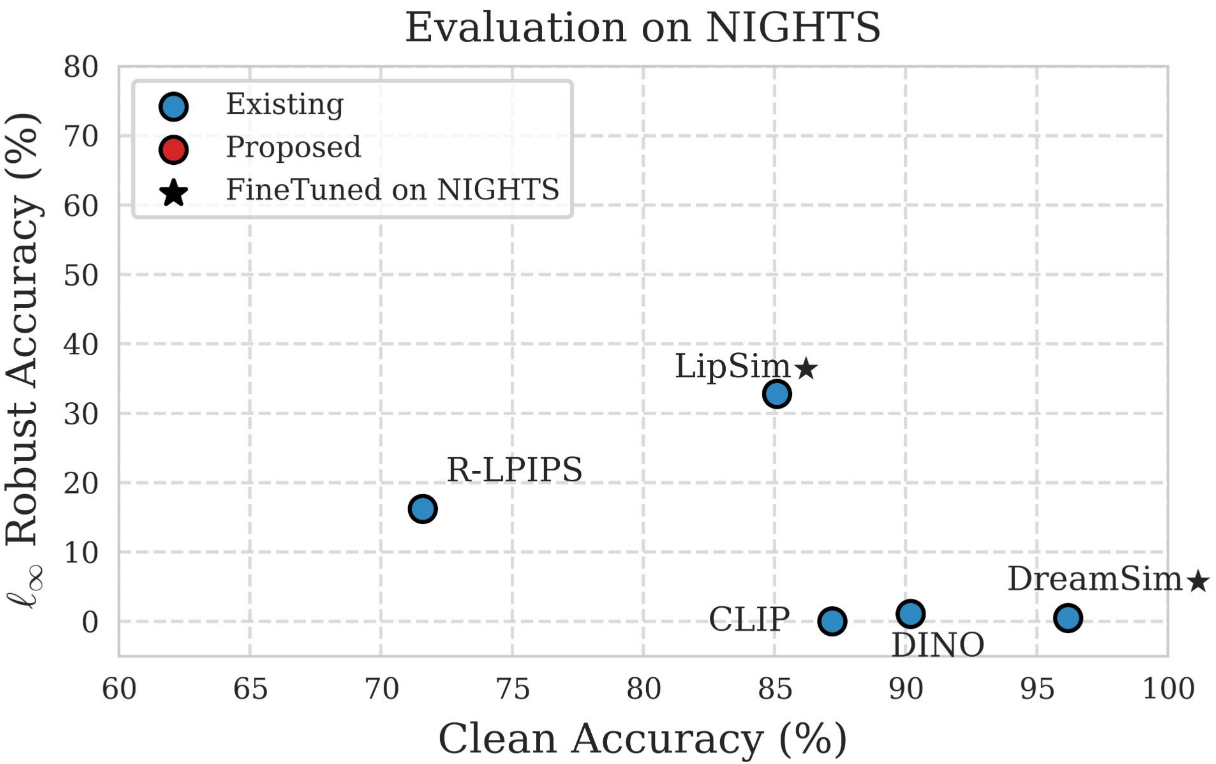
- Fine-tune **only on ImageNet** (without labels), ℓ_∞ radius ε = 4/255.
- Models: CLIP ConvNeXt-B (**R-CLIP$_\text{F}$**) and DINO ViT-B/16 (**R-DINO$_\text{F}$**)

[1] Schlarmann et al., Robust CLIP: Unsupervised Adversarial Fine-Tuning of Vision Embeddings for Robust Large Vision-Language Models, ICML 2024

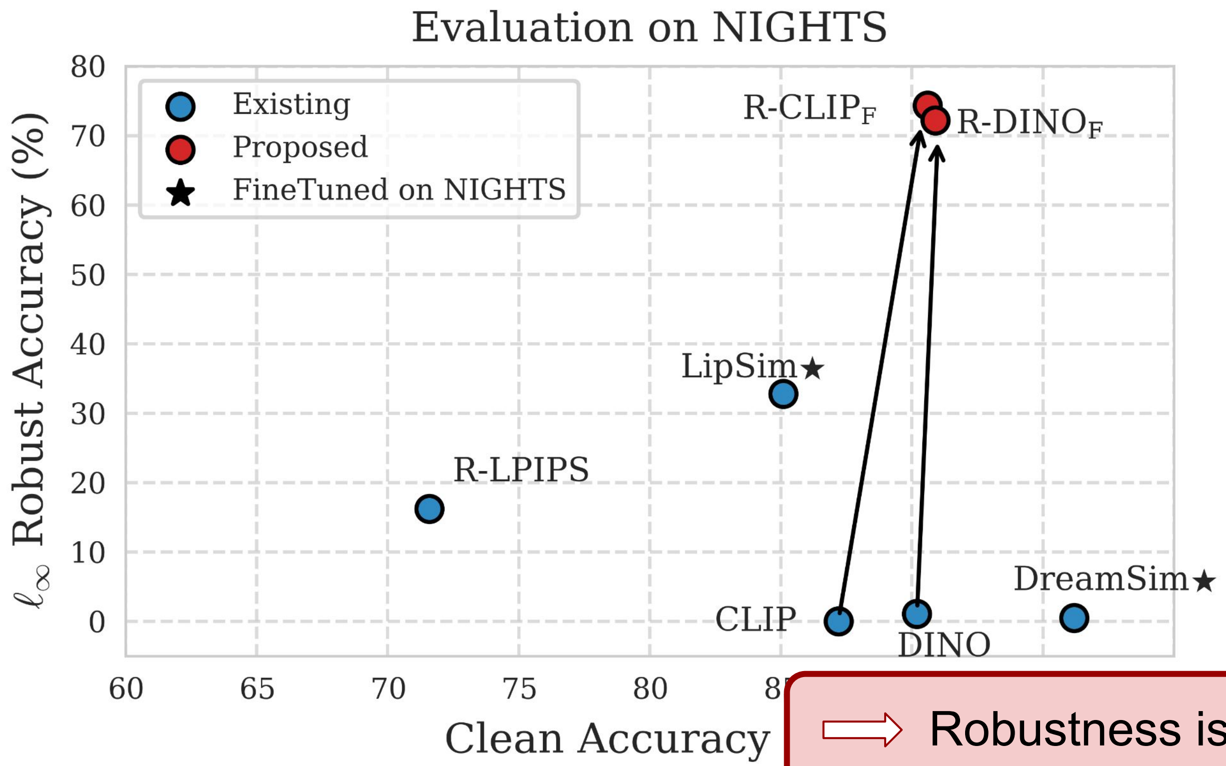
Perceptual Metric Evaluation



Perceptual Metric Evaluation



Perceptual Metric Evaluation



- ➔ **SOTA robustness**
- ➔ **SOTA** zero-shot **clean performance**
- ➔ Clean performance **improves over base model!**

⇨ Robustness is **not** at odds with accuracy!

Content Filtering

- **Goal:** Automatic system that filters unsafe images
- Given a query image, determine whether it is unsafe (**U**) or safe (**S**)
- Can be solved with perceptual metrics via **retrieval**:

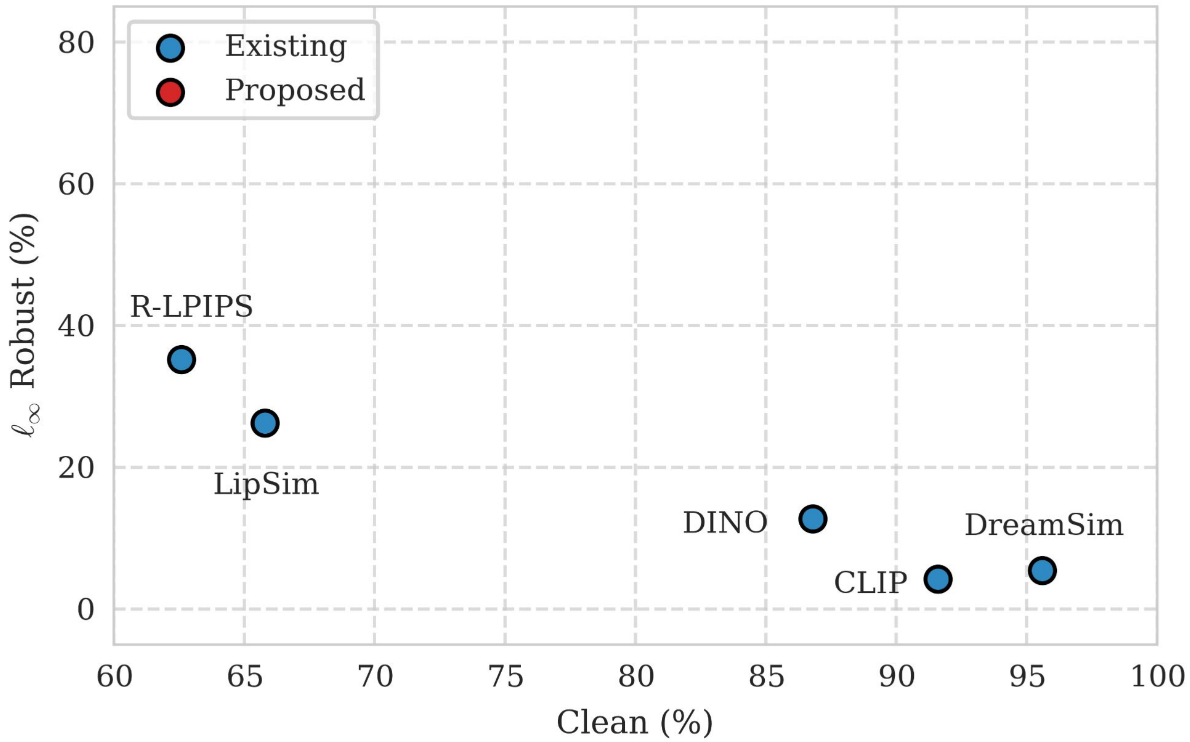
  → is the query image more similar to **U** or **S** images?

Safety critical task ⟹ Robustness is crucial!

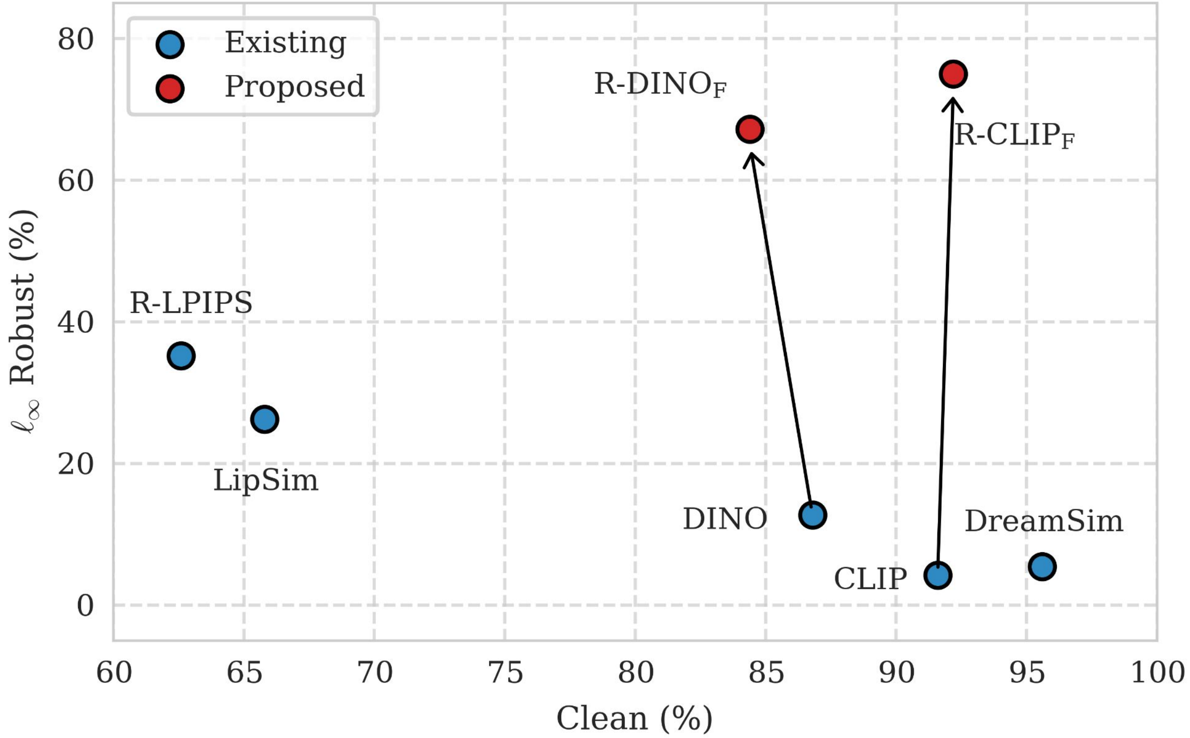
Attack formulation:

- Maximize similarity of unsafe query **x** to small set **Y** of safe images
- No knowledge of retrieval pool required → **realistic scenario**

Robust Content Filtering: Results



Robust Content Filtering: Results



- ➔ **SOTA robustness** in this safety critical task
- ➔ Competitive **clean performance**
- ➔ Clean accuracy improves slightly for CLIP, decreases slightly for DINO

Interpretability

*What images are considered **similar** by the perceptual metrics?*

- **Invert** embedding $\phi(\boldsymbol{x})$
- Solve

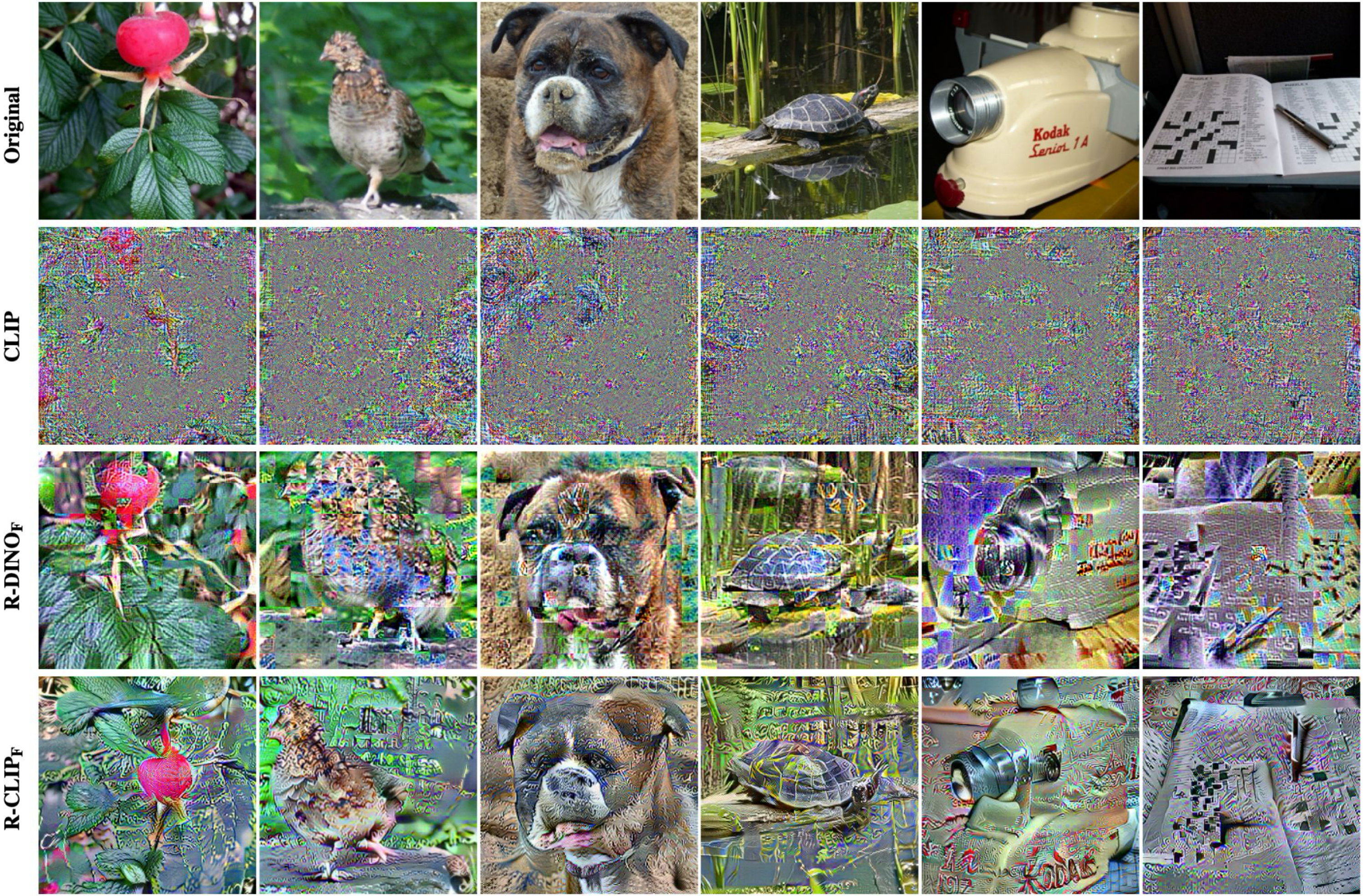
$$\arg\max_{\hat{\boldsymbol{x}} \in [0,1]^d} \ \mathtt{sim}(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \arg\max_{\hat{\boldsymbol{x}} \in [0,1]^d} \ \cos(\phi(\hat{\boldsymbol{x}}), \phi(\boldsymbol{x}))$$

→ Solution is considered similar by the perceptual metric

- Solve via gradient based **optimization**, starting with gray image
  - Produces adversarial noise for clean models
  - **Robust models** are known to have **interpretable gradients**

Interpretability

Interpretability

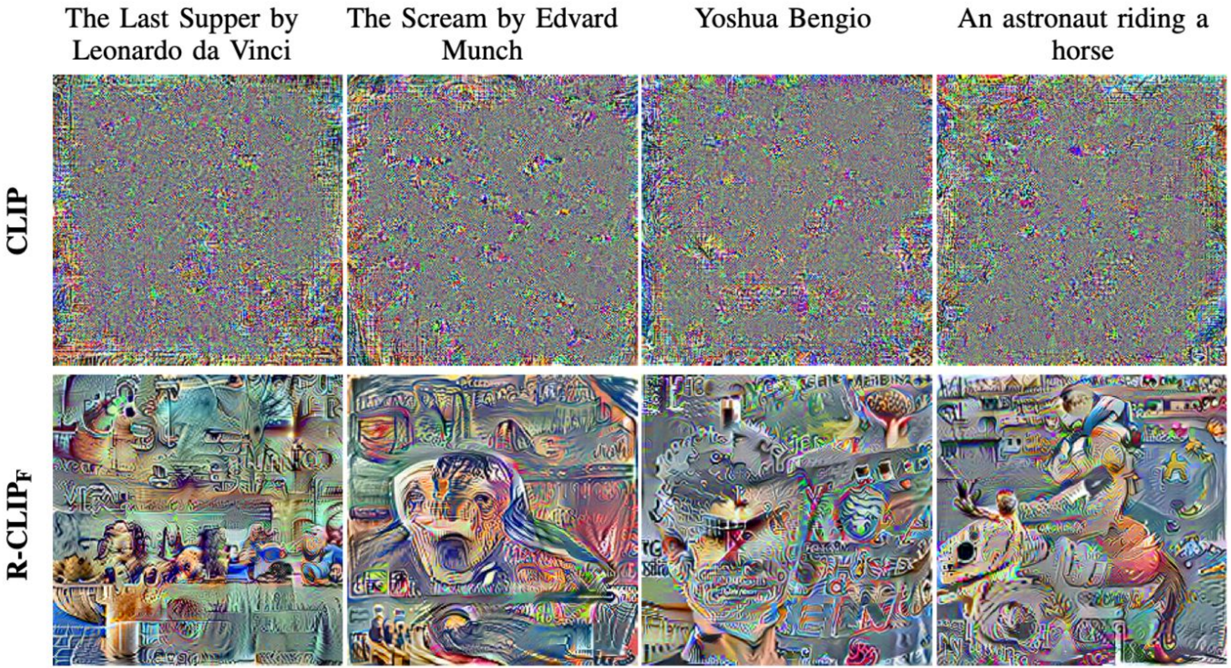


Interpretability

Can also maximize similarity to **text embedding** $\psi(\boldsymbol{t})$:

$$\arg\max_{\boldsymbol{x}\in[0,1]^d} \mathtt{sim}(\boldsymbol{x},\boldsymbol{t}) = \arg\max_{\boldsymbol{x}\in[0,1]^d} \cos(\phi(\boldsymbol{x}),\psi(\boldsymbol{t}))$$

→ extract **concepts** encoded by CLIP



Conclusion

Robust vision encoders yield zero-shot perceptual metrics that

- achieve **SOTA robustness**
- **improve clean performance** over base models
- exhibit **interpretable features**

Code & Models available: