Adversarially Robust CLIP Models Induce Better (Robust) Perceptual Metrics

Francesco Croce [1] **Christian Schlarmann** [2 3] **Naman Deep Singh** [2 3] **Matthias Hein** [2 3]

Abstract

Measuring perceptual similarity is a key tool in computer vision. In recent years perceptual metrics based on features extracted from neural networks with large and diverse training sets, e.g. CLIP, have become popular. At the same time, the metrics extracted from features of neural networks are not adversarially robust. In this paper we show that adversarially robust CLIP models induce *better* and *adversarially robust* perceptual metrics that outperform existing metrics in a zero-shot setting, and further match the performance of state-of-the-art metrics while being robust after fine-tuning. Notably, these perceptual metrics enable adversarially robust NSFW content detection. Finally, the perceptual metrics induced by robust CLIP models have higher interpretability: feature inversion can show which images are considered similar, while text inversion can find what images are associated to a given prompt. This also allows us to visualize the very rich visual concepts learned by a CLIP model, including memorized persons, paintings and complex queries.

1. Introduction

A longstanding goal in computer vision is finding a metric which is able to accurately mimic the human perception of similarity of images. This would benefit multiple tasks such as dataset filtering, image retrieval, copyright infringement discovery, and image quality assessment. While the first approaches to perceptual metrics relied on analyzing statistical properties of the images (Wang et al., 2004), the development of deep learning brought metrics based on internal representations of trained models, among which the most prominent example is the LPIPS distance (Zhang et al., 2018). More recently, the proximity in the embedding space of large foundation models such as CLIP (Radford et al., 2021), DINO (Caron et al., 2021), Masked Autoencoders (MAE) (He et al., 2022) has been shown to effectively capture the semantic similarity of images (Fu et al., 2023). A line of work has further focused on studying the adversarial robustness of perceptual similarity metrics, showing that they are extremely brittle against even imperceptible perturbations (Kettunen et al., 2019; Sjögren et al., 2022; Ghildyal & Liu, 2022; 2023). This might become especially problematic in tasks where an adversary has interest in bypassing automatic similarity checks, e.g. in image attribution, content filtering (Andriushchenko et al., 2022), or "Not Safe for Work" (NSFW) detection in large scale datasets. Recently, Ghazanfari et al. (2023) proposed R-LPIPS, an empirically robust version of LPIPS against ℓ_p adversarial perturbations, and Ghazanfari et al. (2024) introduced LipSim, a first perceptual metric with certified robustness against ℓ_2-bounded perturbations. However, empirical and even more so provable robustness are typically at odds with accuracy.

In our work, we show that recent advancements in robust CLIP models (Mao et al., 2023; Schlarmann et al., 2024) can provide unforeseen benefits for perceptual metrics and their robustness. Surprisingly, these robust models achieve significantly better performance on Two Alternatives Force Choice (2AFC) datasets, showing higher alignment with human judgements, than their clean counterparts or other models like DINO and MAE. Moreover, the induced perceptual metric inherits the robustness of the vision embedding, outperforming the SOTA robust perceptual metrics of LipSim and R-LPIPS by large margins. While previous work on perceptual metrics has solely focused on CLIP models with vision transformers (ViTs) as encoder, we show that the stronger inductive bias of convolutions in ConvNeXt might be particularly effective in this task. Further, our robust CLIP models perform similarly to the original non-robust ones on image-to-image retrieval tasks, while being significantly more robust. This is particularly relevant as it can be translated in making unsafe image detection, via CLIP embedding, robust to malicious attackers, which we experimentally test on a NSFW images dataset. Finally, we illustrate the interpretability of our robust metrics via feature inversion (inverting a given image embedding) and text inversion (generating images from captions).

[1]EPFL, Switzerland [2]Tübingen AI Center, Germany [3]University of Tübingen, Germany. Correspondence to: F. Croce <francesco.croce@epfl.ch>.

Published at ICML 2024 Workshop on Foundation Models in the Wild.

2. Background

Adversarially Robust CLIP Models. CLIP models (Radford et al., 2021) consist of an image encoder $\phi : I \to \mathbb{R}^D$ and a text encoder $\psi : T \to \mathbb{R}^D$, which map different types of data into the same D-dimensional latent space. The embedding of image-text pairs with corresponding semantic meaning are then aligned in the latent space via contrastive learning using large datasets of image-caption pairs. These models attain good results in zero-shot classification performance: the K class names are reformulated as text prompts, e.g. $\boldsymbol{t}_k =$"`A photo of <class k>`" for $k = 1, \dots, K$, and embedded via the text encoder as $\psi(\boldsymbol{t}_k)$. The predicted class for an image $\boldsymbol{x}$ is the one whose text embedding has the highest cosine similarity to the image embedding. As for image classifiers obtained by supervised learning, the zero-shot CLIP classifiers are vulnerable to adversarial perturbations (Fort, 2021; Mao et al., 2023), in particular in the ℓ_p-bounded threat models. Recent works have proposed to extend adversarial training (Madry et al., 2018) to CLIP by fine-tuning the image encoder an existing non-robust CLIP model against ℓ_∞-bounded perturbations: TeCoA (Mao et al., 2023) performs supervised adversarial training on ImageNet, while FARE (Schlarmann et al., 2024) formulates an unsupervised learning problem where one aims at obtaining the same embedding for both clean and adversarially perturbed images (training is on ImageNet images).

CLIP Embedding Induces a Perceptual Metric. To measure the similarity of two images $\boldsymbol{x}_1, \boldsymbol{x}_2 \in I$ it is common to use the cosine similarity of their embedding, i.e. CLIP induces the similarity score

$$\texttt{sim}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \left\langle \frac{\phi(\boldsymbol{x}_1)}{\|\phi(\boldsymbol{x}_1)\|_2}, \frac{\phi(\boldsymbol{x}_2)}{\|\phi(\boldsymbol{x}_2)\|_2} \right\rangle \quad (1)$$

which is used as perceptual metric. This is well-aligned with human perception on the NIGHTS dataset in a 2AFC task (Fu et al., 2023) even in a zero-shot setting, i.e. without fine-tuning on NIGHTS.

2AFC datasets. In Two Alternatives Forced Choice (2AFC) tasks, given a reference image $\boldsymbol{x}_{\text{ref}}$ one has to decide which out of two images $\boldsymbol{x}_1, \boldsymbol{x}_2$ is most similar to the reference image (with ground truth label $y \in \{1, 2\}$). Two popular 2AFC datasets for perceptual metrics are the BAPPS dataset (Zhang et al., 2018), used to tune the LPIPS distance based on features of AlexNet, and the NIGHTS dataset (Fu et al., 2023), used to tune the DreamSim metric. Given a perceptual metric or similarity score, one can formulate this problem as a classification task: with the CLIP embedding we get

$$\texttt{clf}(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_{\text{ref}}) = [\texttt{sim}(\boldsymbol{x}_{\text{ref}}, \boldsymbol{x}_1), \texttt{sim}(\boldsymbol{x}_{\text{ref}}, \boldsymbol{x}_2)] \quad (2)$$

which predicts labels as $\arg\max_{k=1,2} \texttt{clf}(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_{\text{ref}})$. A classifier which performs well on such a 2AFC task is well-aligned with human perception. Given 2AFC training data one can fine-tune the image embedding on this task.

Attacks on perceptual metrics. One can adversarially attack the classifier in Eq. (2) in several ways, applying perturbations either on one of (or both) the test images $\boldsymbol{x}_1, \boldsymbol{x}_2$ or the reference image $\boldsymbol{x}_{\text{ref}}$. We consider the second option more intuitive as it may influence both similarity comparisons, which mimics an attack scenario for image attribution or content filtering. This is also in line with previous work in LipSim (Ghazanfari et al., 2024). The resulting optimization problem for the attack can be formulated as

$$\max_{\|\boldsymbol{\delta}\|_p \leq \epsilon_p} \mathcal{L}(\texttt{clf}(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_{\text{ref}} + \boldsymbol{\delta}), y) \text{ s.th. } \boldsymbol{x}_{\text{ref}} + \boldsymbol{\delta} \in I,$$

which can be solved with PGD-like attacks (Madry et al., 2018; Croce & Hein, 2020) on some classification loss $\mathcal{L}$ e.g. cross-entropy.

3. Evaluation of Perceptual Metrics induced by Robust CLIP Models

In the following we study the effectiveness and robustness of the similarity metrics induced by adversarially robust CLIP models. We consider three CLIP models from the OpenCLIP library (Cherti et al., 2023) with vision encoder using different backbones (ViT-B/32, ViT-B/16, ConvNeXt-B), all pre-trained on LAION-2B (Schuhmann et al., 2022). To get adversarially robust versions, we fine-tune them with FARE and TeCoA on ImageNet (ℓ_∞-threat model with radius $\epsilon_\infty = 4/255$): we indicate them as R-CLIP$_\text{F}$ and R-CLIP$_\text{T}$ respectively. We test adversarial robustness to ℓ_∞-bounded attacks of radius $\epsilon_\infty = 4/255$ and ℓ_2-bounded attacks of size $\epsilon_2 = 3$, as a proxy for unseen threat models. For computing the attacks we use APGD (Croce & Hein, 2020) on the cross-entropy loss for 100 iterations. Details about the experimental setup are in Appendix B, and additional experiments in Appendix C.

3.1. Fine-Tuning for ℓ_∞-Robustness Makes CLIP Models More Aligned with Human Perception

Zero-shot perceptual metrics. Table 1 reports the clean and robust accuracy of CLIP models across different architectures on the test set of NIGHTS. The robust CLIP models achieve significantly higher clean accuracy than their original clean CLIP counterparts, from which they have been fine-tuned. The improvements are consistent across encoder architectures and adversarial fine-tuning schemes (FARE, TeCoA), in the range of 5-6%. This is remarkable as adversarial robustness is typically associated with a loss in performance: we hypothesize that the robustness to imperceptible ℓ_∞-perturbation leads to an emphasis of robust features, which are likely more correlated with higher

Table 1. **Comparison of CLIP and robust CLIP models on NIGHTS.** We report clean and robust accuracy of both zero shot and NIGHTS fine-tuned (either with MLP or LoRA) CLIP models with different vision encoders.

Method	Encoder	clean	ℓ_∞	ℓ_2
Zero-shot CLIP				
Clean	ViT-B/32	85.1	0.0	0.1
R-CLIP$_\text{F}$	ViT-B/32	91.1	71.8	70.6
R-CLIP$_\text{T}$	ViT-B/32	91.0	79.1	**79.7**
Clean	ViT-B/16	85.1	0.0	0.0
R-CLIP$_\text{F}$	ViT-B/16	90.6	71.5	65.5
R-CLIP$_\text{T}$	ViT-B/16	91.9	79.4	77.1
Clean	CnvNxt-B	87.2	0.0	0.0
R-CLIP$_\text{F}$	CnvNxt-B	90.6	74.3	66.1
R-CLIP$_\text{T}$	CnvNxt-B	**92.3**	**81.9**	78.5
MLP Fine-tuned CLIP				
Clean	CnvNxt-B	90.2	0.0	0.0
R-CLIP$_\text{F}$	CnvNxt-B	92.5	78.2	69.0
R-CLIP$_\text{T}$	CnvNxt-B	**94.5**	**84.4**	**79.8**
LoRA Fine-tuned CLIP				
Clean	CnvNxt-B	**95.4**	0.0	0.0
R-CLIP$_\text{F}$	CnvNxt-B	95.3	85.6	81.6
R-CLIP$_\text{T}$	CnvNxt-B	95.0	**87.2**	**84.5**

Table 2. **Comparison to SOTA (robust) perceptual metrics on NIGHTS.** Although the DreamSim Ensemble achieves the best clean performance, our R-CLIP$_\text{T}$ +LoRA model attains the best robustness with high clean performance. FT: represents whether the model is fine-tuned/distilled with NIGHTS. * indicates models not available and robustness could not be evaluated but expected to be similar to DreamSim (Ensemble+LoRA).

Method	Backbone	FT	clean	ℓ_∞	ℓ_2
Perceptual model: **LipSim** (Ghazanfari et al., 2024)					
Pretrained	SLL	✓	86.6	8.6	26.5
Margin$_{0.2}$	SLL	✓	88.5	23.1	46.6
Margin$_{0.5}$	SLL	✓	85.1	32.8	53.1
Perceptual model: **Robust LPIPS** (Ghazanfari et al., 2023)					
R-LPIPS	AlexNet	✗	71.6	16.2	26.9
Perceptual model: **DreamSim** (Fu et al., 2023)					
Ensemble*	ViT-B/16	✗	90.8	-	-
Ensemble + MLP*	ViT-B/16	✓	93.4	-	-
Ensemble + LoRA	ViT-B/16	✓	**96.2**	0.5	0.9
Perceptual model: **Robust CLIP** (ours)					
R-CLIP$_\text{T}$	CnvNxt-B	✗	92.3	81.9	78.5
R-CLIP$_\text{T}$ + MLP	CnvNxt-B	✓	94.5	84.4	79.8
R-CLIP$_\text{T}$ + LoRA	CnvNxt-B	✓	95.0	**87.2**	**84.5**

order semantic concepts. Moreover, the similarity metrics induced by clean CLIP models are, as expected, not adversarially robust. Conversely, using the robust embedding of FARE and TeCoA yields robust perceptual metrics in both ℓ_∞ and ℓ_2-threat models. We observe that the supervised adversarial fine-tuning of TeCoA gives higher robustness in all cases, and typically better clean accuracy, than FARE. Overall, these experiments show that in this case the clean vs robust accuracy trade-off which has been observed in several tasks, see e.g. Tsipras et al. (2019), is even reversed, and adversarial training is beneficial for both clean *and* robust performance. Finally, while Fu et al. (2023) have analyzed models pre-trained on different tasks (CLIP, DINO and MAE), they all share ViTs as backbone for the vision encoders. However, Table 1 illustrates that in our setup the ConvNeXt-B models achieve higher clean and robust accuracy than the two vision transformers of similar size (ViT-B/16, ViT-B/32).

Fine-tuning on the NIGHTS dataset. Fu et al. (2023) also provide a training set in NIGHTS: following their setup we fine-tune our robust ConvNeXt-B R-CLIP$_\text{T}$ on it and report the results in Table 1. With MLP probing, the robust backbone as initialization provides significantly higher clean performance than with a non-robust one. while using LoRA all achieve similar performance (within standard deviation over seeds). Notably, using the adversarially trained backbones allows the similarity metric to retain, and even improve, robustness in the ℓ_p-threat models. While this might be unexpected, we speculate that the fine-tuning, especially with LoRA, allows the models to rely on a subset of a few (the NIGHTS benchmark is of limited difficulty) task-specific features for classification. This means, benefiting from the robust pre-training the relevant features are highly robust, while the further fine-tuning down-weights the importance of non-robust ones, thus leading to the improvement in robustness.

3.2. Comparison to SOTA (Robust) Perceptual Metrics

In Table 2 we compare our R-CLIP$_\text{T}$ (with ConvNeXt-B) to SOTA methods for clean and robust perceptual metrics. Fu et al. (2023) propose the DreamSim-Ensemble which concatenates the features of three ViTs (CLIP, DINO, OpenCLIP) to obtain the features for computing perceptual similarity: this achieved SOTA results on NIGHTS both zero-shot and with fine-tuning, although at increased inference cost. R-CLIP$_\text{T}$ outperforms the DreamSim-Ensemble in both the zero-shot setup and when fine-tuning a task-specific MLP head, while being worse only for LoRA fine-tuning. In the context of robust perceptual metrics, LipSim (Ghazanfari et al., 2024) Pretrained model attains certified ℓ_2-robustness by distilling DreamSim on ImageNet, while the Margin$_{0.2}$ and Margin$_{0.5}$ models are further fine-tuned on NIGHTS. The main goal of LipSim is certified ℓ_2-robustness, but Ghazanfari et al. (2024) also report good performance in empirical robustness. Moreover, Ghazanfari et al. (2023) proposes a robust version of LPIPS trained

Table 3. **(Robust) Accuracy of perceptual metrics on BAPPS.** Our zero-shot R-CLIP$_\text{T}$ is close to or outperforms the baselines. Clean performance is over the entire dataset, while robust accuracy is computed with APGD for 1k images for each split (* LipSim-Pretrained is distilled from DreamSim which in turn is fine-tuned on NIGHTS).

Method	Encoder	FT-Data	clean	ℓ_∞	ℓ_2
Perceptual model: **LipSim** (Ghazanfari et al., 2024)					
Pretrained	SLL	NIGHTS*	74.2	1.1	7.4
Margin$_{0.2}$	SLL	NIGHTS	74.0	5.8	15.1
Margin$_{0.5}$	SLL	NIGHTS	73.1	7.0	12.3
Perceptual model: **Robust LPIPS** (Ghazanfari et al., 2023)					
R-LPIPS	AlexNet	BAPPS	72.8	7.0	12.3
Perceptual model: **DreamSim** (Fu et al., 2023)					
Ensemble + LoRA	ViT-B/16	NIGHTS	73.1	0.0	0.0
Perceptual model: **Robust CLIP** (ours)					
R-CLIP$_\text{T}$	CnvNxt-B	None	74.1	26.8	15.8
R-CLIP$_\text{T}$ + MLP	CnvNxt-B	NIGHTS	74.2	28.5	16.3
R-CLIP$_\text{T}$ + LoRA	CnvNxt-B	NIGHTS	**74.7**	**29.2**	**20.0**

on the BAPPS dataset. The robust CLIP embedding outperforms the baselines: our zero-shot R-CLIP$_\text{T}$ achieves 49.1% and 25.4% higher robust accuracy in ℓ_∞ and ℓ_2 respectively than the LipSim-Margin$_{0.5}$ model, while having even 7.2% better clean performance. Finally, DreamSim does not provide any non-trivial robustness.

3.3. Evaluation on the BAPPS dataset

We further test the effectiveness of our models on BAPPS (Zhang et al., 2018). First, consistently with NIGHTS, robust CLIP encoders improve the clean performance on BAPPS compared to their clean equivalents, as shown in Table 8. Second, we compare R-CLIP$_\text{T}$ to the baselines in Table 3, where we report average performance over the 6 dataset splits (breakdown over splits in Appendix C). Zero-shot and LoRA R-CLIP$_\text{T}$ achieve the same or better performance than the baselines, in particular DreamSim, while having the highest robust accuracy for both threat models. This shows that the (robust) perceptual metric induced by our robust encoders is effective across both 2AFC datasets.

4. Robust Image-to-Image Retrieval

Nearest neighbors retrieval. Perceptual metrics can be used to find the nearest neighbors of a query image $\boldsymbol{x}$ in a pool of retrieval images. To test the adversarial robustness of a metric in this task, we optimize a perturbation $\boldsymbol{\delta}$ ($\epsilon = 4/255$) to maximize the distance between the embedding of $\boldsymbol{x}$ and $\boldsymbol{x} + \boldsymbol{\delta}$, (which does not require access to the retrieval set. For the revisited Oxford and Paris datasets (Radenović et al., 2018), Figure 1 (top plot) shows

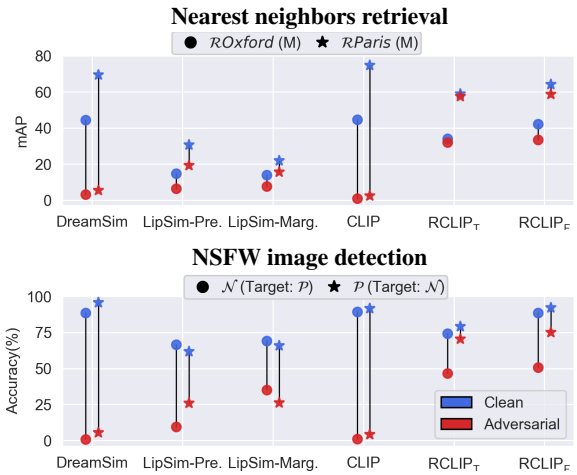

Figure 1. **Robust image-to-image retrieval.** Our R-CLIP$_\text{F}$ (ConvNeXt) retains clean performance (blue) similar to the baselines, while being significantly more robust (red) on both tasks.

that our robust perceptual metrics (zero-shot evaluation) achieve significantly higher robust mean Average Precision (mAP) than the baseline (close to zero for both CLIP and DreamSim), at the cost of a small degradation in clean performance. Further results are provided in Appendix C.6.

Robust NSFW detection. Robust image-to-image retrieval might become particularly relevant when an adversary has an incentive to bypass the automated scanning process, such as filtering NSFW content. To test the different perceptual metrics on this task, we sample 500 images each from a public dataset[1] for the classes 'neutral' ($\mathcal{N}$), 'p*rn' ($\mathcal{P}$) and 's*xy' ($\mathcal{S}$) as retrieval pools. Then, we select test sets of 500 images for the $\mathcal{N}$ and $\mathcal{P}$ classes, disjoint of the retrieval sets. As classification rule, we compute the cosine similarity for each query image to all 1500 retrieval images (3 classes), and select the class of the image with maximal similarity. For adversarial evaluation, the attacks minimize the average similarity between the embedding of the query image and images from the (opposite) target class, using APGD at $\epsilon = 8/255$ (a detailed description of the setup in Appendix C.7). In Figure 1 we show clean and robust accuracies of various perceptual metrics: for both query classes R-CLIP$_\text{F}$ attains clean performance similar to the original CLIP and DreamSim (R-CLIP$_\text{T}$ is in this case slightly worse, possibly due to the supervised fine-tuning). However, CLIP and DreamSim show little adversarial robustness, while R-CLIP$_\text{F}$ preserves 75.0% accuracy under attacks which try to make unsafe images be classified as neutral (i.e. query class $\mathcal{P}$ and target $\mathcal{N}$), which is the most practically relevant scenarios. Detailed results can be found in Tab. 6 in Appendix C.7.

[1] https://huggingface.co/datasets/deepghs/nsfw_detect

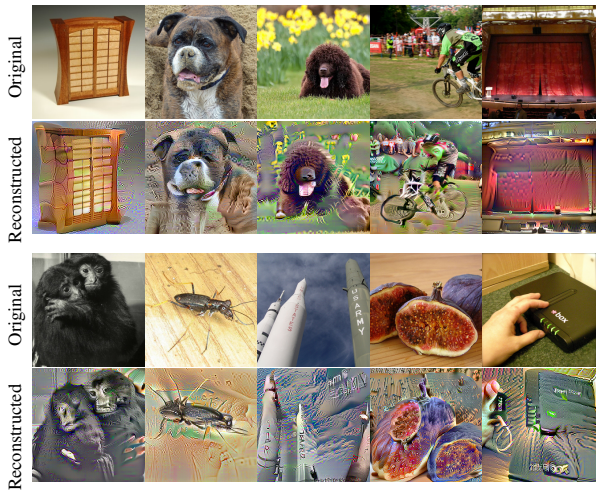

Figure 2. **Feature inversion.** We reconstruct images from their R-CLIP$_\text{T}$ embedding by optimizing a randomly initialized image to maximize similarity in the embedding space. Clear features of the original images are reconstructed.

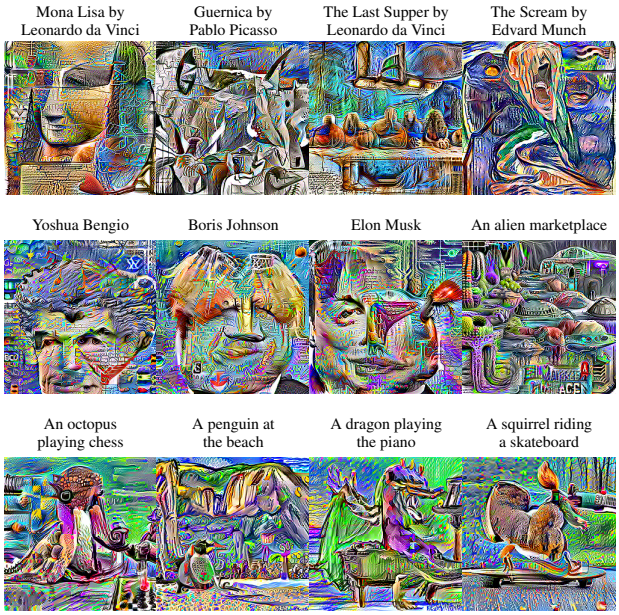

Figure 3. **Text inversion.** We show visual concepts encoded in R-CLIP$_\text{T}$ by optimizing randomly initialized images to match the given text prompts in the embedding space. We are able to extract rich and meaningful visual concepts from R-CLIP$_\text{T}$.

5. Visual Concepts of Robust CLIP Models

Feature inversion. To study which images are considered similar (or identical) by our perceptual metric, we aim at finding images mapped to the same embedding vector. We can formulate such task as finding an image $\hat{\boldsymbol{x}}$ which maximizes the similarity to the embedding $\phi(\boldsymbol{x})$ of a given reference image $\boldsymbol{x}$, i.e. $\arg\max_{\hat{\boldsymbol{x}}\in I} \texttt{sim}(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \arg\max_{\hat{\boldsymbol{x}}\in I} \cos(\phi(\hat{\boldsymbol{x}}), \phi(\boldsymbol{x}))$. As we assume access to the encoded image $\phi(\boldsymbol{x})$ only, it can only be solved approximately. The search space I is the space of all images and thus very large. For regularization, we constrain the optimization to an ℓ_2-ball, and optimize it with APGD initialized at a gray image with small additive uniform noise, see Appendix B for details. Since the formulation is analogous to that of adversarial attacks, if the encoder ϕ was not robust it would not be possible to find meaningful solutions (see also Figure 6). The reconstructed images for R-CLIP$_\text{T}$ (Figure 2) recover quite accurate versions of the original images, where subjects, colors and structure are well approximated, including small details. This is remarkable as the embedding space of ConvNeXt-B is only 640-dimensional, and we use simple constrained optimization.

Text inversion. We can use also the CLIP text encoder ψ to explore which images are associated to a given text prompt by our robust perceptual metric. In practice, we find such images, given a target text $\boldsymbol{t}$, by solving $\arg\max_{\boldsymbol{x}\in I} \texttt{sim}(\boldsymbol{x}, \boldsymbol{t}) = \arg\max_{\boldsymbol{x}\in I} \cos(\phi(\boldsymbol{x}), \psi(\boldsymbol{t}))$ as done for the features inversion experiments. While for non-robust models this would produce mainly noise, as shown in Fig. 6 in Appendix, with our robust model R-CLIP$_\text{T}$ clearly recognizable features of the target text appears, see Fig. 3, although the generated images are highly saturated and show distorted shapes. We show more examples and alternative optimization schemes in Appendix C.4. As feature inversion produces more realistic images than text inversion, we hypothesize that text inversion is more difficult particularly due to the modality gap (Liang et al., 2022). Surprisingly, this simple method can generate complex scenes by closely following the given text prompt (e.g. last row of Figure 3). Also, this shows how CLIP has memorized during training a large number of popular subjects, including paintings and (real or fictional) public figures: then, adversarial fine-tuning emphasizes the reliance of robust features, and allows us to extract such memorized information via optimizing the similarity score.

6. Conclusion

We have shown that fine-tuning CLIP models with adversarial training provides perceptual metrics which significantly better align with human judgement than with clean CLIP models, and achieve SOTA performance for single encoders on 2AFC tasks. At the same time, such metrics inherit the adversarial robustness of the CLIP vision embedding, outperforming existing methods for robust perceptual metrics. Moreover, we illustrate how robust perceptual metrics might be helpful for robust image-to-image retrieval and unsafe content detection. Thanks to these properties, as well as their interpretability, adversarially robust perceptual metrics may find interesting applications in many (safety-critical) tasks.

Acknowledgements

We thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting CS and NDS. We acknowledge support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC number 2064/1, project number 390727645), as well as in the priority program SPP 2298, project number 464101476. We are also thankful for the support of Open Philanthropy and the Center for AI Safety Compute Cluster. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsors.

References

Andriushchenko, M., Li, X. R., Oxholm, G., Gittings, T., Bui, T., Flammarion, N., and Collomosse, J. Aria: Adversarially robust image attribution for content provenance. In *CVPR*, 2022. 1

Araujo, A., Havens, A. J., Delattre, B., Allauzen, A., and Hu, B. A unified algebraic perspective on lipschitz neural networks. In *ICLR*, 2023. 8

Augustin, M., Meinke, A., and Hein, M. Adversarial robustness on in- and out-distribution improves explainability. In *ECCV*, 2020. 8

Boreiko, V., Augustin, M., Croce, F., Berens, P., and Hein, M. Sparse visual counterfactual explanations in image space. In *GCPR*, 2022. 8

Carlini, N. and Wagner, D. A. Adversarial examples are not easily detected: Bypassing ten detection methods. In *AISec*, 2017. 8

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 1, 8, 11, 12

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., and Jitsev, J. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2023. 2, 14, 15, 16, 17

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020. 2, 8, 9, 11

Ding, K., Ma, K., Wang, S., and Simoncelli, E. P. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:2567–2581, 2020. 8

Fort, S. Adversarial examples for the openai clip in its zero-shot classification regime and their semantic generalization, Jan 2021. URL https://stanislavfort.github.io/2021/01/12/OpenAI_CLIP_adversarial_examples.html. 2

Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., and Isola, P. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *NeurIPS*, 2023. 1, 2, 3, 4, 8, 9, 11, 14, 15, 16, 17

Ganz, R. and Elad, M. Clipag: Towards generator-free text-to-image generation. In *WACV*, 2024. 10, 11

Ghazanfari, S., Garg, S., Krishnamurthy, P., Khorrami, F., and Araujo, A. R-LPIPS: An adversarially robust perceptual similarity metric. In *ICML Workshop on New Frontiers in Adversarial Machine Learning*, 2023. 1, 3, 4, 8, 14, 15, 16, 17

Ghazanfari, S., Araujo, A., Krishnamurthy, P., Khorrami, F., and Garg, S. Lipsim: A provably robust perceptual similarity metric. In *ICLR*, 2024. 1, 2, 3, 4, 8, 14, 15, 16, 17

Ghildyal, A. and Liu, F. Shift-tolerant perceptual similarity metric. In *ECCV*, 2022. 1

Ghildyal, A. and Liu, F. Attacking perceptual similarity metrics. *Transactions on Machine Learning Research*, 2023. 1, 8

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 1, 8

Hebart, M. N., Contier, O., Teichmann, L., Rockter, A. H., Zheng, C. Y., Kidder, A., Corriveau, A., Vaziri-Pashkam, M., and Baker, C. I. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12: e82580, feb 2023. ISSN 2050-084X. 11

Kazemi, H., Chegini, A., Geiping, J., Feizi, S., and Goldstein, T. What do we learn from inverting clip models? *arXiv preprint arXiv:2403.02580*, 2024. 10

Kettunen, M., Härkönen, E., and Lehtinen, J. E-lpips: Robust perceptual image similarity via random transformation ensembles. *arXiv:1906.03973*, 2019. 1

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2015. 11

Laidlaw, C. and Feizi, S. Functional adversarial attacks. In *NeurIPS*, 2019. 8

Liang, V. W., Zhang, Y., Kwon, Y., Yeung, S., and Zou, J. Y. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*, 2022. 5

Lin, T., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: common objects in context. In *ECCV (5)*, 2014. 12

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. 2, 8

Mahendran, A. and Vedaldi, A. Understanding deep image representations by inverting them. In *CVPR*, 2015. 8

Mao, C., Geng, S., Yang, J., Wang, X. E., and Vondrick, C. Understanding zero-shot adversarial robustness for large-scale models. In *ICLR*, 2023. 1, 2

Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, 2008. 12, 13

Prashnani, E., Cai, H., Mostofi, Y., and Sen, P. Pieapp: Perceptual image-error assessment through pairwise preference. In *CVPR*, 2018. 8

Radenović, F., Iscen, A., Tolias, G., Avrithis, Y., and Chum, O. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *CVPR*, 2018. 4, 12, 13

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 8, 12, 14, 15, 16

Santurkar, S., Ilyas, A., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Image synthesis with a single (robust) classifier. In *NeurIPS*, 2019. 8

Schlarmann, C. and Hein, M. On the adversarial robustness of multi-modal foundation models. In *ICCV Workshop on Adversarial Robustness In the Real World*, 2023. 9

Schlarmann, C., Singh, N. D., Croce, F., and Hein, M. Robust CLIP: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. *arXiv:2402.12336*, 2024. 1, 2, 8, 11

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., and Jitsev, J. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 2, 12

Sjögren, O., Pihlgren, G. G., Sandin, F., and Liwicki, M. Identifying and mitigating flaws of deep perceptual similarity metrics. *arXiv:2207.02512*, 2022. 1

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. Intriguing properties of neural networks. In *ICLR*, 2014. 8

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. In *ICLR*, 2019. 3

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 1, 8

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 1, 2, 4, 8

Zhao, S., Liu, Z., Lin, J., Zhu, J.-Y., and Han, S. Differentiable augmentation for data-efficient gan training. In *NeurIPS*, 2020. 11

A. Related Work

Perceptual metrics. Low-level pixel-based ℓ_p-metrics and structural similarity SSIM (Wang et al., 2004) do not capture well higher-order semantic similarity. These are outperformed by metrics based on features extracted from neural networks trained on ImageNet, such as LPIPS (Zhang et al., 2018) PIE-APP (Prashnani et al., 2018) and DISTS (Ding et al., 2020). More recently, it has been shown that metrics induced by the features extracted by models trained on larger datasets and via self-supervised training, like CLIP (Radford et al., 2021), DINO (Caron et al., 2021) or MAE (He et al., 2022), are well aligned with human perception regarding semantic similarity (Fu et al., 2023). DreamSim (Fu et al., 2023) is a fine-tuned ensemble of three of these models which shows the best alignment with human preferences on the NIGHTS dataset.

Adversarial robustness of perceptual metrics. Virtually all vision tasks tackled via neural networks are vulnerable to adversarial examples (Szegedy et al., 2014), and attacks in several threat models exist (Carlini & Wagner, 2017; Croce & Hein, 2020; Laidlaw & Feizi, 2019). The main empirical defense which works across different vision tasks is adversarial training (Madry et al., 2018). However, the price of having a robust model is typically a drop in performance. Not surprisingly, perceptual metrics, including LPIPS and DreamSim, are also not robust to adversarial perturbations (Ghazanfari et al., 2023; 2024; Ghildyal & Liu, 2023). In order to get a robust version R-LPIPS of the popular LPIPS metric, Ghazanfari et al. (2023) perform adversarial training on the 2AFC fine-tuning task of the Berkeley-Adobe Perceptual Patch Similarity dataset (BAPPS) (Zhang et al., 2018). LipSim (Ghazanfari et al., 2024) distills from DreamSim (Fu et al., 2023) a 1-Lipschitz network, and then fine-tunes it on the NIGHTS (Fu et al., 2023) dataset to achieve certified adversarial robustness.

Generative properties of adversarially robust models. Feature inversion, i.e. finding an image which matches given features at the output layer, can be used to understand the inner workings of a network. However, it often yields highly distorted images without much semantic content (Mahendran & Vedaldi, 2015). At the same time, adversarially robust models suffer significantly less from this problem, and can be used to generate semantically meaningful images when maximizing the probability of a specific class (Santurkar et al., 2019). This can be even exploited to generate visual counterfactuals (instance-specific explanations) for modern image classifiers (Augustin et al., 2020; Boreiko et al., 2022).

B. Experimental Details

We here provide details about the models and setup used in the experiments.

B.1. Models and Evaluation

CLIP models. We use the vision encoders from the OpenCLIP library, and in particular those of CLIP models pre-trained on LAION-2B. The specific model identifiers are listed in Tab. 4. We fine-tune with FARE and TeCoA for 2 epochs for the ℓ_∞-threat model with radius $\epsilon_\infty = 4/255$, following the scheme in Schlarmann et al. (2024). For fine-tuning on NIGHTS we follow the scheme of Fu et al. (2023) (for the ConvNeXt-B encoder we apply LoRA on the fully connected layers of the MLPs).

Table 4. **Model keys from OpenCLIP of different pre-trained encoders.**

Encoder	Identifier key
ViT-B/32	`CLIP-ViT-B-32-laion2B-s34B-b79K`
ViT-B/16	`CLIP-ViT-B-16-laion2B-s34B-b88K`
ConvNeXt-Base	`CLIP-convnext_base_w-laion2B-s13B-b82K-augreg`

Baselines. We use the original DreamSim models, including three single encoders (OpenCLIP, CLIP, DINO) and the corresponding ensemble, all fine-tuned on NIGHTS, as publicly available.[2] The LipSim metric uses a Semi-Definite program based Lipschitz Layers (SLL) convolutional network from Araujo et al. (2023) as the backbone. In the evaluation we use the original LipSim models.[3] Finally, for R-LPIPS we use the model[4] trained for ℓ_∞-robustness on the reference image (on the BAPPS dataset).

[2] https://github.com/ssundaram21/dreamsim
[3] https://github.com/SaraGhazanfari/lipsim
[4] https://github.com/SaraGhazanfari/R-LPIPS

Evaluation. For all models we use image resolution of 224x224, including the clean CLIP with ConvNeXt-B encoder which was pre-trained at 256x256. While the NIGHTS dataset already contains high resolution images, which are then resized and cropped (the exact pre-processing depends on the model), the BAPPS dataset is typically used at 64x64 resolution, then we upsample the images to 224x224. The adversarial perturbations are similarly applied on the 224x224 images.

B.2. Visual Concepts

The images to be optimized are always initialized grey (all pixels 0.5) with small additive uniform noise in $[-8/255, 8/255]$.

Feature inversions. For feature inversions (Fig. 2) we set the ℓ_2 radius to 100 and run APGD (Croce & Hein, 2020) for 500 iterations with initial step-size 200.

Text inversions. For text inversions (Figs. 7, 8) we set the ℓ_2 radius to 200 and run APGD for 100 iterations with initial step-sizes 10 (small) and 400 (large).

C. Additional Experiments

In the following we provide additional evaluations of our robust CLIP models and the induced perceptual metrics.

C.1. Zero-shot performance on the NIGHTS dataset

Table 7 reports the clean and robust accuracy of clean and robust CLIP models across different architectures on the test set of the NIGHTS dataset on both its ImageNet and non-ImageNet splits,[5] and the average over the entire set. The advantage of the robust encoders can be observed on both test splits. In line with Fu et al. (2023), we also observe that the larger robust ViT-L/14 models of Schlarmann & Hein (2023) perform worse than our smaller ViT-B networks.

C.2. Detailed Results on 2AFC Datasets

Fine-tuning different encoders on NIGHTS. To complement the results of Table 1, in Table 7 we show the performance of the perceptual metrics obtained by fine-tuning the CLIP models with different backbones (ViT-B/32, ViT-B/16 and ConvNeXt-B) and pre-training (clean, FARE, TeCoA) on the NIGHTS dataset. When fine-tuning an MLP on top of the frozen encoder, the robust backbones preserve, across architectures, the advantage in both clean and robust accuracy they show in the zero-shot setup compared to the clean CLIP. With LoRA, all backbones achieve similar clean performance, but the metrics based on robust CLIP encoders are the only ones with non-trivial robustness.

Comparison to single DreamSim models. Additionally, we report in Table 7 the results of the variants of DreamSim which use a single ViT as encoder (Fu et al., 2023) and are fine-tuned on NIGHTS with LoRA. We observe that our R-CLIP$_\text{T}$ with ConvNeXt-B backbone plus MLP matches or improves the performance of 2 out of 3 DreamSim models, although it keeps the encoder unchanged (zero-shot setting). Moreover, several of ours model fine-tuned with LoRA perform on par with the best DreamSim model (that is OpenCLIP ViT-B/32 pre-trained on LAION-400M, while our CLIP models have been pre-trained on LAION-2B). Finally, the single DreamSim models come with robust accuracy close to zero in both threat models, unlike our metrics.

Varying perturbation radius. We test our R-CLIP$_\text{T}$ (zero-shot and with LoRA fine-tuning, ConvNeXt-B backbone) and the most robust LipSim model when varying the perturbation radius for both ℓ_∞ and ℓ_2-threat models. Figure 4 shows the clean and robust accuracy of each model on the NIGHTS test set. We observe that our models attain higher robust accuracy than LipSim across radii, while reaching zero at sufficiently large values.

Detailed comparison on BAPPS. Table 8 shows the breakdown of the clean performance of the various perceptual metrics over the 6 splits of BAPPS (the entire validation set is used for this). Consistently with NIGHTS, the adversarially trained CLIP encoders provide a significant improvement compared to their clean counterparts. Also, our models used in the zero-shot setup outperform the DreamSim ones, and are on par with the LipSim metrics (both fine-tuned on NIGHTS). Fine-tuning R-CLIP$_\text{F}$ and R-CLIP$_\text{T}$ on NIGHTS yields some small but consistent increase in clean accuracy. Finally, Table 9 reports the robust accuracy for all metrics in both ℓ_p-threat models: similar to NIGHTS, R-CLIP$_\text{T}$ attains outperforms the existing methods. Interestingly, in this case the R-CLIP$_\text{T}$ with ViT-B/32 backbone show better results than the other

[5] the ImageNet split contains images generated from classes included in ImageNet, see Fu et al. (2023) for details

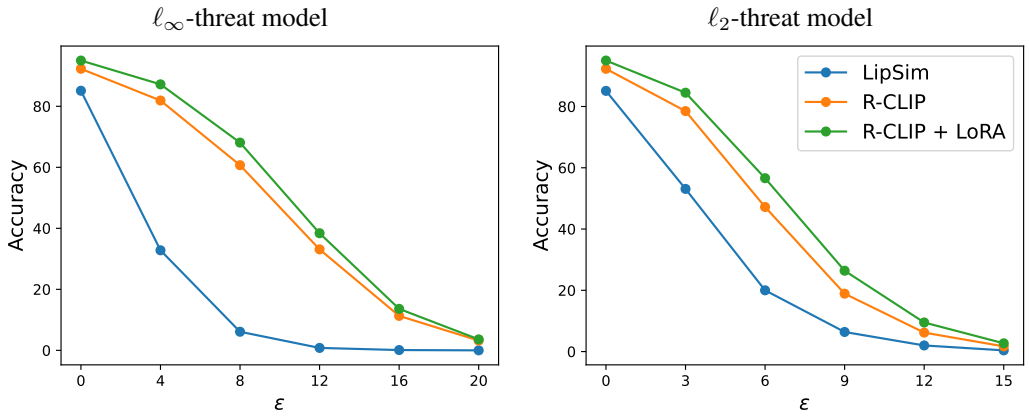

Figure 4. **Robustness at different perturbation radii for NIGHTS.** We show the robust accuracy for our R-CLIP$_\text{T}$ (ConvNeXt-B backbone, zero-shot and fine-tuned with LoRA) and LipSim when varying the perturbation radii, for both ℓ_∞ (left) and ℓ_2 (right) bounded attacks. R-CLIP$_\text{T}$ models outperform LipSim across perturbation sizes.

Figure 5. **Feature inversion variants**. Varying the random seeds for the initialization recovers, when using R-CLIP$_\text{T}$, multiple images for the same target feature. These are sometimes horizontally flipped but preserve the original semantic content.

architectures.

C.3. Feature Inversion

We show in Figure 5 the effect of varying initialization on the optimization results when doing feature inversion: interestingly, the reconstructed images differ mainly in non-semantic aspects, such as small translations or horizontal flip. These examples show that the perceptual metric given by the robust CLIP seems to capture well the semantic content of the images and ignore other aspects, as human would do, which are not prioritized when judging similarity.

C.4. Text Inversion

Small step size. We test the effect of using a smaller initial step size in APGD than the one which gives the images shown in Figure 3 (see Appendix B.2 for details). A comparison of the resulting images with both large and small step size are shown in Figure 7 and Figure 8. The small step size produces more fine-grained visualizations, but in some cases no features are generated. Using the large step size yields features more reliably, although the generated images are highly saturated and show distorted shapes.

Optimization with multiple augmentations. It has been observed that integrating augmentations into the text inversion process improves the quality of generated images (Ganz & Elad, 2024; Kazemi et al., 2024). We test whether it also helps

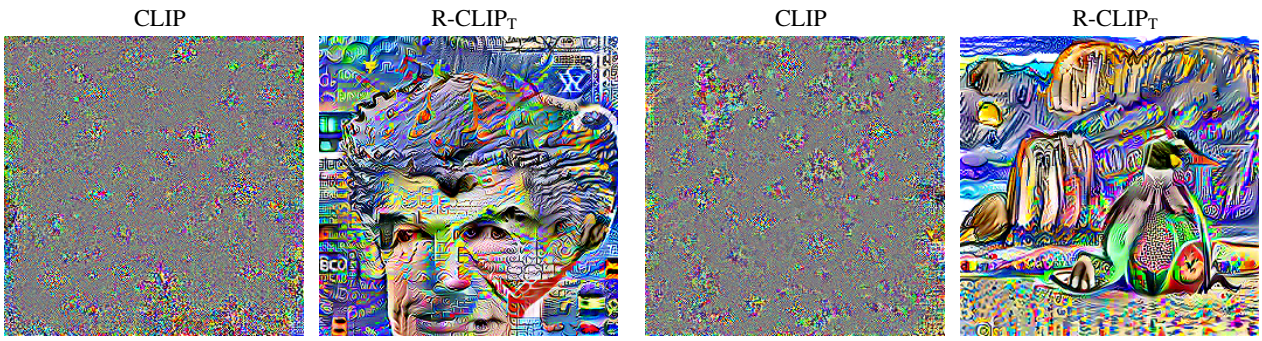

Figure 6. **Robust CLIP makes feature inversion possible.** Starting from a grey image we maximize the cosine similarity to the text embedding of a text query once for CLIP and once for R-CLIP$_\text{T}$ (Left: "Yoshua Bengio", Right: "A penguin at the beach"). Our robust CLIP model generates images which show that both concepts are captured/memorized, whereas the clean CLIP model produces only adversarial noise.

in our setting. To this end, we run the optimization following the setup of Ganz & Elad (2024), i.e. using the Adam optimizer (Kingma & Ba, 2015) for 1000 iterations with initial step-size 0.1 and independently augmenting 32 views of the image via diffaugment (Zhao et al., 2020) with color, translation and cutout augmentations. The results are shown in Figure 9. We observe that this procedure leads to significantly less high-frequency artifacts. However, we are more interested in the interpretability of the perceptual metric than in text-to-image generation. As we achieve lower loss values with the unaugmented process, we show that in the main paper.

Using a clean CLIP model. In Figure 6 we show the effect of performing text inversion with the original CLIP encoder. For both text prompts, the optimization (with APGD) finds only noise-like images, showing that using a robust model is crucial for the interpretability of the induced perceptual metric.

C.5. Evaluation of Models on Different Tasks

Robustness on ImageNet and zero-shot classification. It is interesting to see if the performance on the perceptual metric task is correlated with other properties of CLIP models like zero-shot classification. Therefore, we test the original CLIP and R-CLIP$_\text{T}$ with ConvNeXt-B architecture on ImageNet (note that the adversarial fine-tuning is done on this dataset) and zero-shot image classification for 13 datasets, similar to (Schlarmann et al., 2024). In Table 10, we report the clean and robust accuracy for ℓ_∞ threat model at perturbation strengths of $2/255$ and $4/255$. Robustness is computed with the first two attacks of AutoAttack (Croce & Hein, 2020), i.e. APGD on the cross-entropy and targeted DLR loss. As expected, the two epoch adversarial fine-tuning results in a decay in clean performance with a significant robustness gain across architectures. Thus, while we see for zero-shot classification the usual robustness-accuracy trade-off, this does not hold for the 2AFC-task of the induced perceptual metric. Exploring this difference is an interesting future research direction.

Performance on THINGS dataset. In Table 11, we show how different perceptual models perform on THINGS (Hebart et al., 2023) dataset which contains image triplets with categorical variations and classifies the odd-one-out. Our R-CLIP$_\text{F}$ performs the best followed by the clean CLIP, whereas all fine-tuned models are notably worse. This finding is in line with Fu et al. (2023), who drew the similar conclusion that fine-tuning on NIGHTS degrades the performance on this task.

C.6. Image-to-image retrieval

Next we look at the task of nearest neighbor retrieval, following the setup of Caron et al. (2021): given a query image $\boldsymbol{x} \in I$, perceptual metrics can be used to find its nearest neighbors in a pool of retrieval images, i.e. those with highest similarity. To generate adversarial attacks on this task, we add ℓ_∞-bounded perturbations to the query image to distort its embedding according to the image encoder ϕ. Formally, we maximize the normalized embedding distance between the output of image encoder of the two images using squared ℓ_2-distance. The resulting optimization problem for the attack can be formulated as

$$\max_{\|\boldsymbol{\delta}\|_p \leq \epsilon_p} \left\| \frac{\phi(\boldsymbol{x}+\boldsymbol{\delta})}{\|\phi(\boldsymbol{x}+\boldsymbol{\delta})\|_2} - \frac{\phi(\boldsymbol{x})}{\|\phi(\boldsymbol{x})\|_2} \right\|_2^2 \quad \text{s. th.} \quad \boldsymbol{x}+\boldsymbol{\delta} \in I, \tag{3}$$

which is equivalent to minimizing the cosine similarity $\texttt{sim}(\boldsymbol{x}+\boldsymbol{\delta}, \boldsymbol{x})$, see Eq. (1). Note this can be seen an an untargeted attacks, and does not require access to the retrieval set.

Quantitative results for image retrieval. We consider the Medium (M) and Hard (H) splits of the revisited Oxford ($\mathcal{R}$Oxford) and Paris ($\mathcal{R}$Paris) image retrieval datasets (Philbin et al., 2008; Radenović et al., 2018), whose specific task is to find the images portraying the same landmark as the query image, and report the Mean Average Precision (mAP). For both $\mathcal{R}$Oxford and $\mathcal{R}$Paris the number of query images is 70 whereas the data pool for retrieval contains 5k and 6.3k images respectively, and we always evaluate with image size 224×224 with single-scale, unlike Caron et al. (2021) who evaluate $\mathcal{R}$Paris at a higher resolution with multi-scale view. For adversarial evaluation, we use ℓ_∞ radius to $4/255$, and APGD (100 iterations) as optimizer. Table 5 shows that the clean CLIP models (ViT-B/16 and ConvNeXt backbones) yield the best clean performance, and the DreamSim is ensemble slightly worse, but are completely non-robust under attack. Conversely, the LipSim models are marginally robust, but have low clean performance. Our R-CLIP$_\text{T}$ and R-CLIP$_\text{F}$ models attain clean mAP very close to the DreamSim models, but, unlike the baselines, do not suffer significant degradation against adversarial attacks. R-CLIP$_\text{F}$ (unsupervised) models attain much better clean performance, in some case close to that of the original CLIP, in comparison to the R-CLIP$_\text{T}$ (supervised) models.

Qualitative results for image retrieval. We further evaluate the perceptual metrics for image retrieval on MS-COCO (Lin et al., 2014) dataset. In Figs. 10 and 11, for each query image we first show its nearest neighbour, among a random subset of 15k images from the training data, as identified by the similarity score induced by different models. Then, we apply on the query image an adversarial perturbation, generated at ℓ_∞-radius of $2/255$ using 50 iterations of APGD, which aims at maximizing the squared ℓ_2-distance between the clean and the adversarial embedding (see Eq. (3)). In the 'Adv.' row, we show the nearest neighbour assigned to the perturbed query. From this random set of images we can confer that R-CLIP$_\text{T}$ performs on this task similarly DreamSim for clean inputs. Moreover, we find R-CLIP$_\text{T}$ less susceptible to the adversarial perturbations than DreamSim and LipSim, as it retrieves on average the most relevant image to the query image.

C.7. Robust NSFW classification

Image-to-image retrieval can be used for content filtering. For example, detecting "Not Safe for Work" (NSFW) images is a pressing problem as modern training datasets are often scrapped from the web (Radford et al., 2021; Schuhmann et al., 2022), and unsafe content needs to be discarded from such datasets. Naturally, safe-guarding filtering models against malicious users becomes an important concern.

In this section, we expand on the experimental setup for the robust NSFW detection task presented in Section 4. We sample 500 images each from a public dataset[6] for the classes 'neutral' ($\mathcal{N}$), 'p*rn' ($\mathcal{P}$) and 's*xy' ($\mathcal{S}$). All images include mostly humans, the images associated in class $\mathcal{N}$ are neutral (safe), while class $\mathcal{P}$ includes extreme NSFW cases, and $\mathcal{S}$ is in the middle (unclear). These 500 images each form the retrieval pool for each class. We keep the intermediate $\mathcal{S}$ class to represent cases on which the detection model is uncertain. For computing the adversarial perturbations, we sample a set Y of 16 images belonging the target class (but not included in the retrieval set) and minimize the average distance of their normalized embeddings to that of query image $\boldsymbol{x}$. This yields the optimization problem

$$\min_{\|\boldsymbol{\delta}\|_p \leq \epsilon_p} \sum_{\boldsymbol{y} \in Y} \left\| \frac{\phi(\boldsymbol{x}+\boldsymbol{\delta})}{\|\phi(\boldsymbol{x}+\boldsymbol{\delta})\|_2} - \frac{\phi(\boldsymbol{y})}{\|\phi(\boldsymbol{y})\|_2} \right\|_2^2 \quad \text{s. th.} \quad \boldsymbol{x}+\boldsymbol{\delta} \in I, \tag{4}$$

which is optimized with 200 iterations of APGD at ℓ_∞-radius of $\epsilon = 8/255$. We notice that Eq. (4) can be seen as the targeted version of Eq. (3).

In Table 6 we provide the detailed results of detection accuracy of different perceptual models, which is summarized in Figure 1. The original CLIP model performs best in clean performance on neutral images (class $\mathcal{N}$), whereas the DreamSim ensemble on unsafe queries (class $\mathcal{P}$). LipSim models have instead relatively low clean accuracy. On this task, R-CLIP$_\text{T}$ gets significantly worse accuracy than CLIP, possibly to the supervised fine-tuning which degrades the performance on image distributions far from that ImageNet. R-CLIP$_\text{F}$, which relies on unsupervised fine-tuning, performs in fact on par with CLIP, while having achieving robust accuracy of 50.6% and 75.0% on query from class $\mathcal{N}$ and $\mathcal{P}$ respectively. Conversely, the performance of both CLIP and the DreamSim ensemble degrades below 6%, and only the DreamSim DINO model has non-trivial robustness.

[6] https://huggingface.co/datasets/deepghs/nsfw_detect

Table 5. **Quantitative robust image-to-image retrieval.** We show the clean and robust mAP (mean Average Precision) on both Medium (M) and Hard (H) sets of datasets proposed by Philbin et al. (2008) for the image retrieval task as formulated in Radenović et al. (2018). The best performing model in each column is highlighted.

		$\mathcal{R}$Oxford				$\mathcal{R}$Paris			
		clean		$\ell_\infty(4/255)$		clean		$\ell_\infty(4/255)$	
Method	Encoder	M	H	M	H	M	H	M	H
Perceptual model: **LipSim**									
Pretrained	SLL	14.7	2.1	6.4	1.6	30.6	9.0	19.2	6.2
Margin$_{0.5}$	SLL	13.9	2.1	7.7	1.6	21.2	7.3	15.6	5.3
Perceptual model: **DreamSim**									
OpenClip	ViT-B/32	39.5	12.2	0.9	0.4	64.2	37.6	3.0	1.6
DINO	ViT-B/16	31.1	8.0	1.2	0.6	59.0	30.2	5.4	2.2
Ensemble	-	44.5	15.1	0.8	0.4	69.4	43.3	3.2	1.4
Perceptual model: **CLIP**									
CLIP	ViT-B/16	**47.2**	**16.0**	1.0	0.5	74.3	51.8	2.8	1.9
R-CLIP$_\text{T}$	ViT-B/16	31.6	8.0	27.9	7.7	53.1	26.1	49.8	23.4
R-CLIP$_\text{F}$	ViT-B/16	37.0	10.0	29.2	9.2	59.7	33.2	55.9	28.2
CLIP	ConvNeXt-B	44.7	14.4	0.9	0.5	**74.6**	**52.0**	2.4	1.7
R-CLIP$_\text{T}$	ConvNeXt-B	34.1	10.3	32.1	9.3	58.8	32.2	57.4	30.2
R-CLIP$_\text{F}$	ConvNeXt-B	42.2	13.0	**33.4**	**10.1**	64.1	37.2	**58.6**	**32.2**

Table 6. **Robust NSFW detection.** We consider both scenarios: *(i)* when query images are from $\mathcal{N}$ and target is $\mathcal{P}$ and, *(ii)* when query images are from $\mathcal{P}$ and target is from $\mathcal{N}$. We report for both cases, the fraction of points allocated to each of the 3 classes with and without (clean) adversarial attack.

		Query: $\mathcal{N}$ Target: $\mathcal{P}$						Query: $\mathcal{P}$ Target: $\mathcal{N}$					
		clean			$\ell_\infty(8/255)$			clean			$\ell_\infty(8/255)$		
Method	Encoder	$\mathcal{N}$	$\mathcal{S}$	$\mathcal{P}$	$\mathcal{N}$	$\mathcal{S}$	$\mathcal{P}$	$\mathcal{N}$	$\mathcal{S}$	$\mathcal{P}$	$\mathcal{N}$	$\mathcal{S}$	$\mathcal{P}$
Perceptual model: **LipSim**													
Pretrained	SLL	66.4	20.8	12.8	9.6	32.0	58.4	5.6	32.8	61.6	61.6	12.6	25.8
Margin$_{0.5}$	SLL	69.2	16.0	14.8	35.0	15.2	49.8	21.6	12.6	65.8	50.2	23.6	26.2
Perceptual model: **DreamSim**													
DINO	ViT-B/16	72.2	14.0	13.8	5.0	11.4	83.6	0.6	6.6	92.8	63.8	16.6	19.6
Ensemble	-	88.6	7.8	3.6	0.8	3.2	96.0	0.2	4.8	**95.6**	84.0	10.6	5.4
Perceptual model: **CLIP**													
CLIP	ConvNeXt-B	**89.4**	6.6	4.0	1.0	9.8	89.2	0.2	8.2	91.6	89.0	6.8	4.2
R-CLIP$_\text{T}$	ConvNeXt-B	74.2	9.0	16.8	46.6	21.0	32.4	9.8	11.2	79.0	8.2	21.4	70.4
R-CLIP$_\text{F}$	ConvNeXt-B	88.6	4.2	7.2	**50.6**	15.2	34.2	1.2	6.6	92.2	18.6	6.4	**75.0**

Table 7. **Comparison of perceptual metrics on NIGHTS dataset.** For each model we show clean and robust accuracy computed with APGD_{CE} with 100 iterations.

		ImageNet			non-ImageNet			Average		
Model	Backbone	clean	ℓ_∞	ℓ_2	clean	ℓ_∞	ℓ_2	clean	ℓ_∞	ℓ_2
Perceptual model: **CLIP** (Radford et al., 2021; Cherti et al., 2023) / **Robust CLIP** (ours)										
Clean	ViT-B/32	85.7	0.0	0.1	84.3	0.0	0.0	85.1	0.0	0.1
R-CLIP$_\text{F}$	ViT-B/32	92.3	74.7	72.2	89.4	67.8	68.5	91.1	71.8	70.6
R-CLIP$_\text{T}$	ViT-B/32	91.9	81.1	80.8	89.8	76.5	78.3	91.0	79.1	79.7
Clean	ViT-B/16	86.3	0.0	0.0	83.5	0.0	0.0	85.1	0.0	0.0
R-CLIP$_\text{F}$	ViT-B/16	91.1	74.7	67.1	90.0	67.2	63.3	90.6	71.5	65.5
R-CLIP$_\text{T}$	ViT-B/16	93.1	82.5	78.6	90.3	75.2	75.1	91.9	79.4	77.1
Clean	ConvNeXt-B	88.0	0.0	0.0	86.1	0.0	0.0	87.2	0.0	0.0
R-CLIP$_\text{F}$	ConvNeXt-B	91.5	78.3	68.0	89.3	69.1	63.6	90.6	74.3	66.1
R-CLIP$_\text{T}$	ConvNeXt-B	92.9	83.6	79.2	91.4	79.7	77.5	92.3	81.9	78.5
Clean	ViT-L/14	83.2	0.0	0.0	79.8	0.0	0.0	81.7	0.0	0.0
R-CLIP$_\text{F}$	ViT-L/14	88.0	69.0	52.4	86.1	60.7	51.9	87.2	65.4	52.2
R-CLIP$_\text{T}$	ViT-L/14	90.4	79.3	74.6	87.5	69.2	68.8	89.1	74.9	72.1
Perceptual model: **LipSim** (Ghazanfari et al., 2024)										
Pretrained	SLL	87.0	8.0	26.5	86.1	9.5	26.6	86.6	8.6	26.5
Margin$_{0.2}$	SLL	90.2	22.9	46.9	86.2	23.5	46.2	88.5	23.1	46.6
Margin$_{0.5}$	SLL	86.1	33.4	55.1	83.9	31.9	50.3	85.1	32.8	53.1
Perceptual model: **R-LPIPS** (Ghazanfari et al., 2023)										
R-LPIPS	AlexNet	72.4	15.7	27.8	70.5	17.0	25.8	71.6	16.2	26.9
Perceptual model: **DreamSim** (Fu et al., 2023)										
OpenCLIP	ViT-B/32	96.4	1.6	2.9	94.1	2.0	3.8	95.4	1.8	3.3
CLIP	ViT-B/32	94.1	0.1	0.3	93.6	0.1	0.4	93.9	0.1	0.3
DINO	ViT-B/16	94.6	3.1	5.8	94.4	4.2	6.8	94.5	3.6	6.2
Ensemble	-	96.6	0.4	0.7	95.5	0.6	1.3	96.2	0.5	0.9
Perceptual model: **MLP Fine-tuned CLIP** (ours)										
Clean	ViT-B/32	91.1	0.0	0.0	87.5	0.0	0.0	89.5	0.0	0.0
R-CLIP$_\text{F}$	ViT-B/32	94.1	77.5	74.4	92.7	69.6	70.1	93.5	74.1	72.6
R-CLIP$_\text{T}$	ViT-B/32	93.6	84.5	83.2	90.5	79.9	80.8	92.3	82.6	82.2
Clean	ViT-B/16	90.2	0.0	0.0	87.1	0.0	0.0	88.9	0.0	0.0
R-CLIP$_\text{F}$	ViT-B/16	93.2	79.6	70.4	92.1	70.9	65.8	92.7	75.9	68.4
R-CLIP$_\text{T}$	ViT-B/16	95.5	84.9	82.0	91.3	78.4	75.9	93.7	82.1	79.4
Clean	ConvNeXt-B	91.2	0.0	0.0	89.0	0.0	0.0	90.2	0.0	0.0
R-CLIP$_\text{F}$	ConvNeXt-B	93.0	80.7	70.8	92.0	74.8	66.7	92.5	78.2	69.0
R-CLIP$_\text{T}$	ConvNeXt-B	95.1	87.0	81.0	93.6	80.8	78.3	94.5	84.4	79.8
Perceptual model: **LoRA Fine-tuned CLIP** (ours)										
Clean	ViT-B/32	95.6	0.3	1.0	93.7	0.4	0.9	94.8	0.5	0.9
R-CLIP$_\text{F}$	ViT-B/32	96.1	83.2	82.1	94.4	77.5	79.8	95.3	80.8	81.1
R-CLIP$_\text{T}$	ViT-B/32	94.9	82.8	83.2	93.5	78.2	80.8	94.3	80.8	82.2
Clean	ViT-B/16	95.2	0.0	0.0	93.6	0.0	0.0	94.5	0.0	0.0
R-CLIP$_\text{F}$	ViT-B/16	95.8	83.0	78.8	95.5	78.0	78.3	95.7	80.9	78.6
R-CLIP$_\text{T}$	ViT-B/16	95.0	84.1	82.6	94.0	78.0	79.4	94.6	81.5	81.2
Clean	ConvNeXt-B	95.5	0.0	0.0	95.3	0.0	0.0	95.4	0.0	0.0
R-CLIP$_\text{F}$	ConvNeXt-B	96.0	87.9	82.2	94.5	82.5	80.7	95.3	85.6	81.6
R-CLIP$_\text{T}$	ConvNeXt-B	95.6	89.3	85.2	94.3	84.3	83.7	95.0	87.2	84.5

Table 8. **Detailed comparison of perceptual metrics on different splits of the BAPPS dataset.** We report the clean accuracy of each model on the 6 splits of the BAPPS dataset, together with their mean. Robust CLIP encoders provide consistent improvements across splits.

model	Backbone	cnn	color	deblur	frameint.	superr.	trad.	mean
Perceptual model: **CLIP** (Radford et al., 2021; Cherti et al., 2023) / **Robust CLIP** (ours)								
Clean	ViT-B/32	83.1	61.1	58.6	63.0	70.3	78.2	69.1
R-CLIP$_\text{F}$	ViT-B/32	86.7	70.8	65.0	67.3	76.1	78.7	74.1
R-CLIP$_\text{T}$	ViT-B/32	86.5	71.5	65.0	67.4	76.6	77.7	74.1
Clean	ViT-B/16	81.9	60.3	55.5	65.2	68.9	77.9	68.3
R-CLIP$_\text{F}$	ViT-B/16	87.3	70.1	64.9	68.2	75.9	78.3	74.1
R-CLIP$_\text{T}$	ViT-B/16	86.4	71.2	64.9	67.2	76.4	77.7	74.0
Clean	ConvNeXt-B	82.3	60.0	55.3	65.9	67.3	78.5	68.2
R-CLIP$_\text{F}$	ConvNeXt-B	86.5	70.5	64.8	67.7	75.0	79.3	74.0
R-CLIP$_\text{T}$	ConvNeXt-B	86.6	71.4	64.9	67.5	75.8	78.6	74.1
Perceptual model: **LipSim** (Ghazanfari et al., 2024)								
Pretrained	SLL	86.4	69.9	65.6	66.7	76.8	79.5	74.2
Margin$_{0.2}$	SLL	85.2	71.9	64.7	66.8	77.2	77.9	74.0
Margin$_{0.5}$	SLL	83.6	71.1	64.1	66.0	76.8	77.0	73.1
Perceptual model: **R-LPIPS** (Ghazanfari et al., 2023)								
R-LPIPS	AlexNet	87.5	67.4	63.7	66.5	76.1	75.7	72.8
Perceptual model: **DreamSim** (Fu et al., 2023)								
OpenCLIP	ViT-B/32	86.4	67.0	63.0	65.6	74.7	81.7	73.1
CLIP	ViT-B/32	83.9	63.3	58.2	63.3	70.0	79.1	69.6
DINO	ViT-B/16	85.7	67.5	62.7	67.3	73.3	80.1	72.8
Ensemble	-	86.7	67.6	62.4	66.3	74.3	81.3	73.1
Perceptual model: **MLP Fine-tuned CLIP** (ours)								
Clean	ViT-B/32	84.4	63.1	58.6	63.9	70.1	78.8	69.8
R-CLIP$_\text{F}$	ViT-B/32	86.8	71.4	65.1	67.0	76.4	78.2	74.2
R-CLIP$_\text{T}$	ViT-B/32	86.3	72.1	64.7	67.2	76.6	77.2	74.0
Clean	ViT-B/16	83.2	62.5	55.8	65.5	69.4	78.4	69.1
R-CLIP$_\text{F}$	ViT-B/16	87.3	71.4	65.2	68.2	76.1	78.3	74.4
R-CLIP$_\text{T}$	ViT-B/16	86.5	71.2	64.9	67.1	76.5	77.7	74.0
Clean	ConvNeXt-B	82.4	62.2	55.7	65.6	67.2	78.9	68.7
R-CLIP$_\text{F}$	ConvNeXt-B	86.8	70.4	64.9	67.2	75.6	78.6	73.9
R-CLIP$_\text{T}$	ConvNeXt-B	86.8	72.0	65.1	67.4	75.6	78.3	74.2
Perceptual model: **LoRA Fine-tuned CLIP** (ours)								
Clean	ViT-B/32	86.0	67.9	60.9	64.5	72.7	81.3	72.2
R-CLIP$_\text{F}$	ViT-B/32	86.8	73.2	65.2	68.2	77.2	79.8	75.1
R-CLIP$_\text{T}$	ViT-B/32	86.2	72.6	65.3	66.9	76.9	77.4	74.2
Clean	ViT-B/16	85.5	66.9	57.4	64.0	72.8	81.0	71.3
R-CLIP$_\text{F}$	ViT-B/16	86.6	72.1	65.4	66.4	76.9	79.7	74.5
R-CLIP$_\text{T}$	ViT-B/16	86.3	72.3	65.0	67.7	76.7	78.5	74.4
Clean	ConvNeXt-B	85.6	65.8	58.1	65.3	72.2	80.4	71.2
R-CLIP$_\text{F}$	ConvNeXt-B	87.5	72.6	65.0	67.2	76.3	80.6	74.9
R-CLIP$_\text{T}$	ConvNeXt-B	87.3	72.4	65.3	67.3	76.0	79.9	74.7

Table 9. **Comparison of perceptual models on the BAPPS dataset with APGD_{CE} 100x1.** We report both ℓ_∞ and ℓ_2 robust accuracy evaluated at radii $4/255$ and 3 respectively for 1k samples on every split of the BAPPS dataset, and their mean.

		cnn		color		deblur		frameinterp.		superres		trad.		mean	
model	Backbone	ℓ_∞	ℓ_2	ℓ_∞	ℓ_2	ℓ_∞	ℓ_2	ℓ_∞	ℓ_2	ℓ_∞	ℓ_2	ℓ_∞	ℓ_2	ℓ_∞	ℓ_2
Perceptual model: **CLIP** (Radford et al., 2021; Cherti et al., 2023) / **Robust CLIP** (ours)															
Clean	ViT-B/32	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.2	0.2	0.0	0.0	0.0	0.0
R-CLIP$_\text{F}$	ViT-B/32	36.1	31.7	17.3	14.5	7.9	4.3	7.0	4.8	17.5	14.5	35.9	29.4	20.3	16.5
R-CLIP$_\text{T}$	ViT-B/32	46.3	44.0	27.9	28.0	17.4	14.7	11.6	10.6	30.3	25.9	41.7	38.7	29.2	27.0
Clean	ViT-B/16	0.0	0.1	0.0	0.0	0.1	0.1	0.0	0.0	0.3	0.3	0.0	0.0	0.1	0.1
R-CLIP$_\text{F}$	ViT-B/16	34.1	16.7	15.2	6.8	6.8	1.9	5.0	2.0	16.4	6.0	38.4	18.6	19.3	8.7
R-CLIP$_\text{T}$	ViT-B/16	45.4	33.7	24.2	19.5	16.0	8.2	10.2	7.0	27.0	16.8	44.1	32.3	27.8	19.6
Clean	ConvNeXt-B	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.2	0.2	0.0	0.0	0.0	0.0
R-CLIP$_\text{F}$	ConvNeXt-B	34.1	11.4	14.9	5.2	7.0	0.7	5.5	1.7	15.9	3.5	36.4	13.6	19.0	6.0
R-CLIP$_\text{T}$	ConvNeXt-B	45.2	28.1	21.3	13.0	13.6	4.30	11.3	5.5	26.7	15.0	42.8	29.1	26.8	15.8
Perceptual model: **LipSim** (Ghazanfari et al., 2024)															
Pretrained	SLL	2.8	15.5	0.3	3.2	0.1	1.0	0.0	0.2	0.9	4.9	2.4	19.3	1.1	7.4
Margin$_{0.2}$	SLL	9.7	25.6	3.8	11.3	0.1	2.2	0.0	1.1	1.7	7.3	9.8	28.8	4.2	12.7
Margin$_{0.5}$	SLL	14.0	28.8	8.3	20.5	0.2	2.2	0.0	1.0	1.7	6.0	10.7	31.9	5.8	15.1
Perceptual model: **R-LPIPS** (Ghazanfari et al., 2023)															
R-LPIPS	AlexNet	20.8	31.3	8.8	13.9	0.2	0.8	1.2	2.3	3.0	5.6	8.3	20.0	7.0	12.3
Perceptual model: **DreamSim** (Fu et al., 2023)															
OpenCLIP	ViT-B/32	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.2	0.2	0.0	0.0	0.0	0.0
CLIP	ViT-B/32	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.2	0.2	0.0	0.0	0.0	0.0
DINO	ViT-B/16	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.2	0.2	0.1	0.0	0.0	0.1
Ensemble	-	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.2	0.2	0.0	0.0	0.0	0.0
Perceptual model: **MLP Fine-tuned CLIP** (ours)															
Clean	ViT-B/32	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.2	0.2	0.0	0.0	0.0	0.0
R-CLIP$_\text{F}$	ViT-B/32	38.7	32.6	21.2	17.5	9.1	4.7	7.1	5.5	18.5	13.8	38.5	28.9	22.2	17.2
R-CLIP$_\text{T}$	ViT-B/32	47.2	43.5	32.3	31.3	17.4	13.7	11.8	10.4	30.3	24.5	41.9	37.8	30.1	26.9
Clean	ViT-B/16	0.0	0.1	0.0	0.0	0.1	0.2	0.0	0.0	0.3	0.6	0.0	0.0	0.1	0.2
R-CLIP$_\text{F}$	ViT-B/16	37.7	18.2	20.5	10.6	8.1	1.6	6.2	2.3	18.0	6.0	38.7	16.8	21.5	9.3
R-CLIP$_\text{T}$	ViT-B/16	46.3	33.4	29.0	21.4	16.7	7.2	10.6	6.8	28.3	16.3	43.7	31.8	29.1	19.5
Clean	ConvNeXt-B	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.2	0.2	0.0	0.0	0.0	0.0
R-CLIP$_\text{F}$	ConvNeXt-B	38.6	14.1	20.7	7.7	8.6	0.6	6.7	2.5	17.4	3.7	38.2	14.2	21.7	7.1
R-CLIP$_\text{T}$	ConvNeXt-B	47.4	28.7	28.0	18.4	14.2	4.1	11.5	6.0	26.6	12.6	43.2	28.0	28.5	16.3
Perceptual model: **LoRA Fine-tuned CLIP** (ours)															
Clean	ViT-B/32	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.2	0.2	0.0	0.0	0.0	0.0
R-CLIP$_\text{F}$	ViT-B/32	38.1	35.4	24.7	23.7	7.1	5.1	7.6	7.0	18.6	17.4	36.9	30.6	22.2	19.9
R-CLIP$_\text{T}$	ViT-B/32	39.3	42.1	27.0	28.0	6.8	7.7	5.6	6.7	15.6	18.5	35.3	35.6	21.6	23.1
Clean	ViT-B/16	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.2	0.2	0.0	0.0	0.0	0.0
R-CLIP$_\text{F}$	ViT-B/16	43.3	31.9	28.9	22.8	8.5	3.6	8.3	4.6	21.5	9.5	40.7	27.9	25.2	16.7
R-CLIP$_\text{T}$	ViT-B/16	45.7	41.7	34.8	31.2	11.4	7.7	10.2	8.5	24.2	19.5	43.9	36.0	28.4	24.1
Clean	ConvNeXt-B	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.2	0.2	0.0	0.0	0.0	0.0
R-CLIP$_\text{F}$	ConvNeXt-B	48.2	35.8	32.8	24.3	12.5	3.7	10.7	5.4	27.4	16.2	49.5	34.9	30.2	20.1
R-CLIP$_\text{T}$	ConvNeXt-B	48.1	35.1	28.7	23.1	13.4	6.1	10.3	5.6	26.7	14.5	47.7	35.8	29.2	20.0

Table 10. **ImageNet and zero-shot downstream datasets evaluation.** We show the clean and robust accuracy for the original CLIP models and the robust fine-tuned ones with TeCoA on ImageNet. Moreover, we show the same statistics averaged over 13 zero-shot datasets.

		ImageNet			Avg. other datasets		
Method	Encoder	clean	$\ell_\infty(2/255)$	$\ell_\infty(4/255)$	clean	$\ell_\infty(2/255)$	$\ell_\infty(4/255)$
CLIP	ViT-B/32	66.1	0.0	0.0	70.4	0.0	0.0
R-CLIP$_\text{T}$	ViT-B/32	58.3	41.5	25.8	46.8	34.5	23.3
CLIP	ViT-B/16	70.1	0.0	0.0	71.7	0.0	0.0
R-CLIP$_\text{T}$	ViT-B/16	64.0	47.9	31.9	51.5	38.4	26.4
CLIP	ConvNeXt-B	71.8	0.0	0.0	71.6	0.0	0.0
R-CLIP$_\text{T}$	ConvNeXt-B	67.1	51.7	35.3	56.2	44.1	31.8

Table 11. **Comparison of perceptual metrics on THINGS dataset.** We report clean accuracy on the odd-one-out task of THINGS. In this case fine-tuning on NIGHTS is typically detrimental for clean performance (* LipSim-Pretrained is distilled from DreamSim which in turn is fine-tuned on NIGHTS).

Method	Backbone	Source	Fine-tuning dataset	clean acc.
Perceptual model: **CLIP**				
CLIP	ConvNeXt-B	(Cherti et al., 2023)	None	50.7
Perceptual model: **Robust CLIP**				
R-CLIP$_\text{F}$	ConvNeXt-B	ours	None	51.2
R-CLIP$_\text{T}$	ConvNeXt-B	ours	None	48.1
Perceptual model: **LipSim**				
Pretrained	SLL	(Ghazanfari et al., 2024)	NIGHTS*	43.6
Margin$_{0.2}$	SLL	(Ghazanfari et al., 2024)	NIGHTS	41.3
Margin$_{0.5}$	SLL	(Ghazanfari et al., 2024)	NIGHTS	38.5
Perceptual model: **Robust LPIPS**				
R-LPIPS	AlexNet	(Ghazanfari et al., 2023)	BAPPS	38.3
Perceptual model: **Fine-tuned DreamSim**				
OpenCLIP	ViT-B/32	(Fu et al., 2023)	NIGHTS	47.9
CLIP	ViT-B/32	(Fu et al., 2023)	NIGHTS	49.6
DINO	ViT-B/16	(Fu et al., 2023)	NIGHTS	44.3
Ensemble	-	(Fu et al., 2023)	NIGHTS	47.5
Perceptual model: **Fine-tuned Robust CLIP**				
R-CLIP$_\text{T}$ + MLP	ConvNeXt-B	ours	NIGHTS	47.6
R-CLIP$_\text{T}$ + LoRA	ConvNeXt-B	ours	NIGHTS	49.9

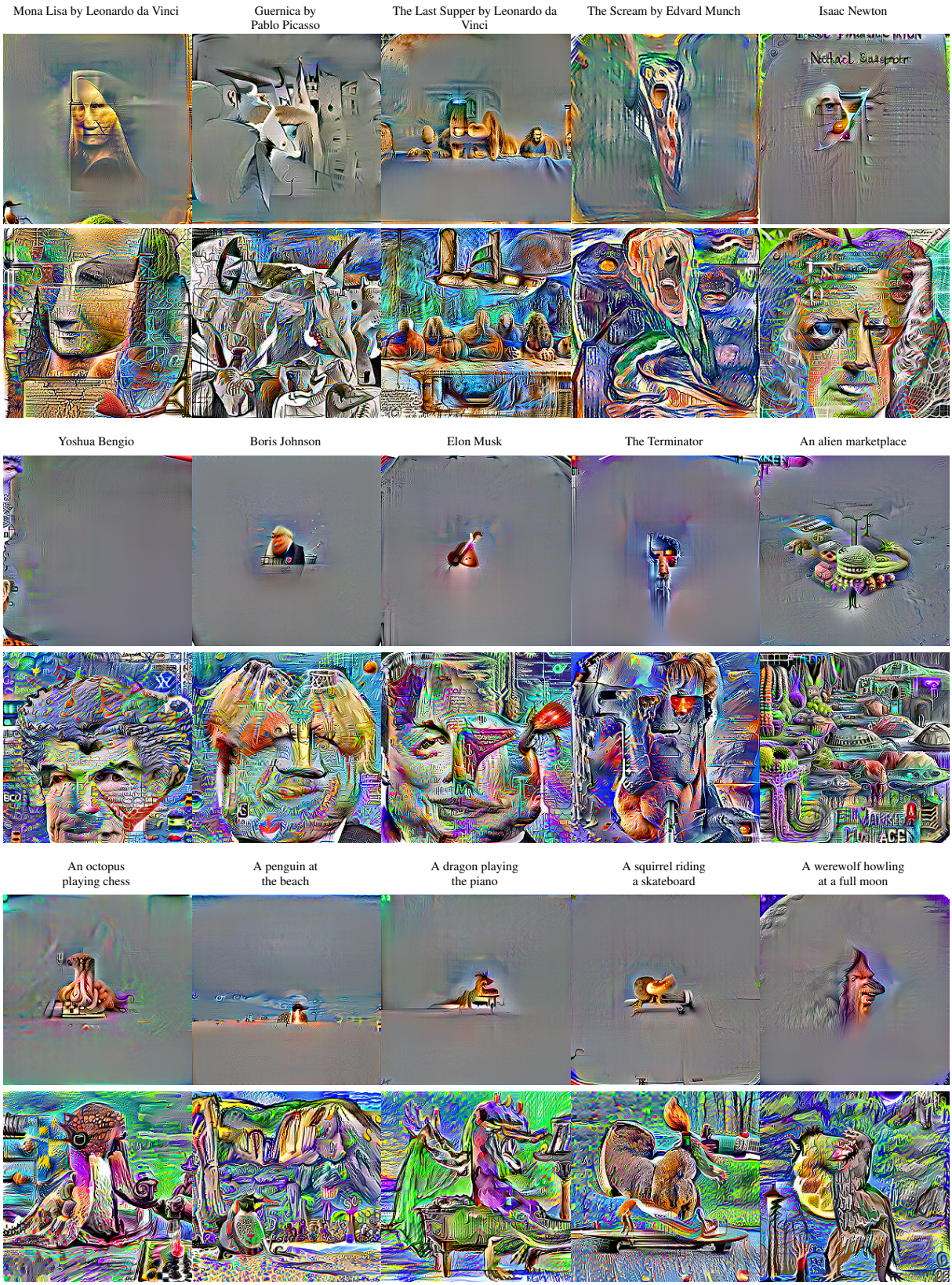

Figure 7. **Text inversion.** We show visual concepts encoded in R-CLIP$_\text{T}$ by optimizing randomly initialized images to match the given text prompts in the embedding space. Small initial step-size and large initial step-size are considered in the first and second rows respectively. We are able to extract rich and meaningful visual concepts from R-CLIP$_\text{T}$.

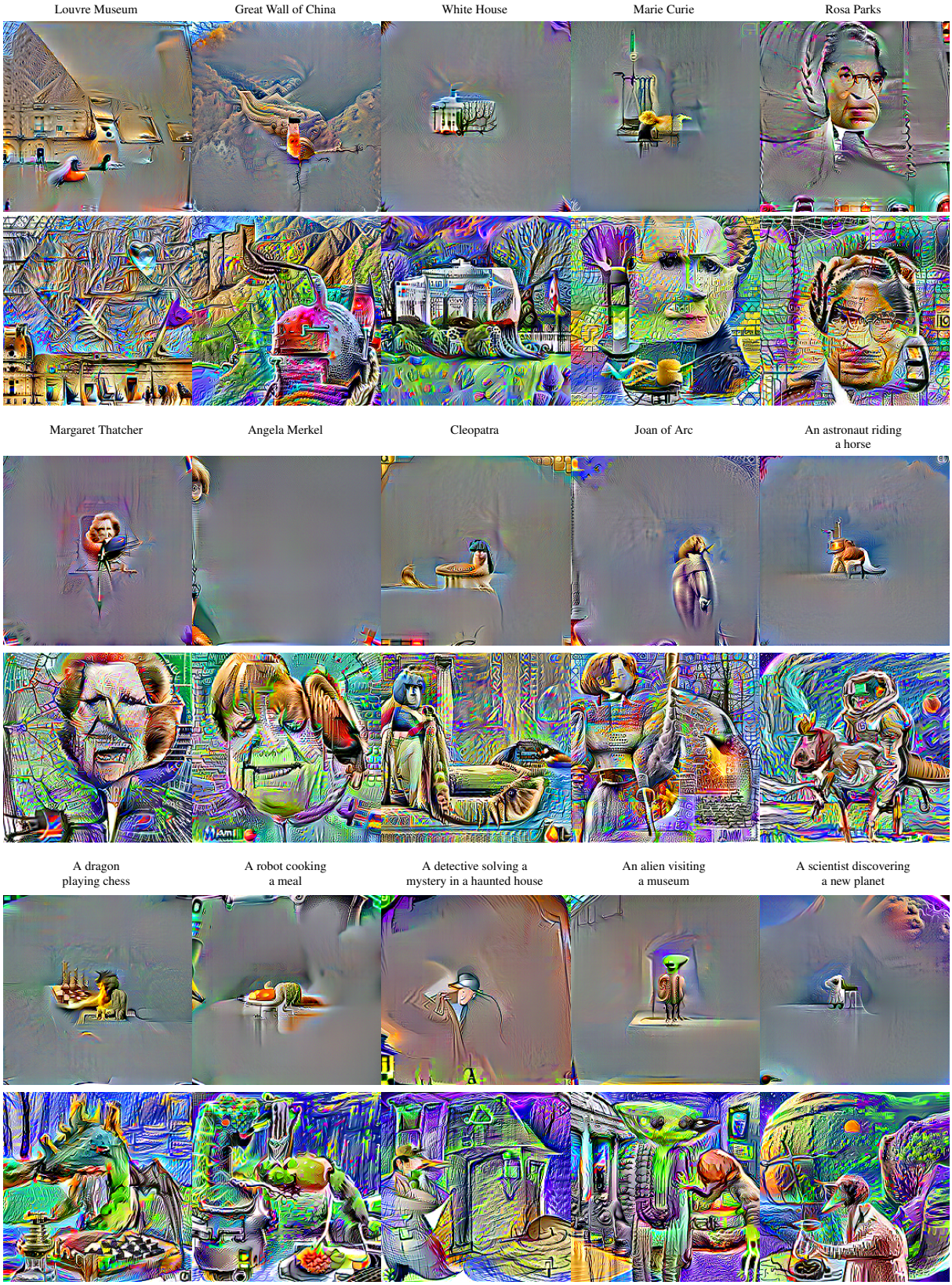

Figure 8. **Text inversion.** We show additional text inversion examples for R-CLIP$_{\text{T}}$. Small initial step-size and large initial step-size are considered in the first and second rows respectively.

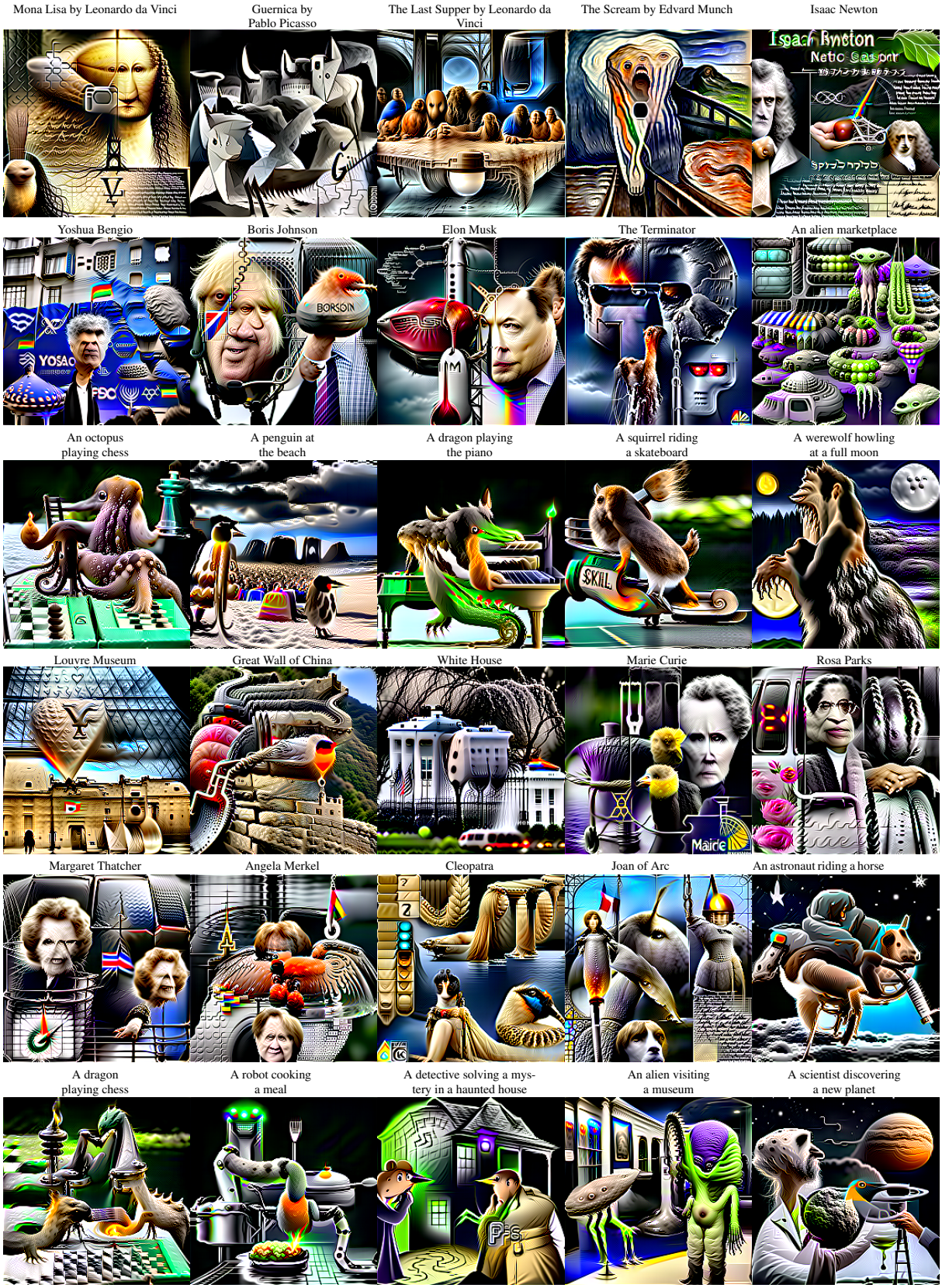

Figure 9. **Text inversion variant.** The text inversions in this figure are created using an augmentation procedure in the optimization process as described in Sec. C.4.

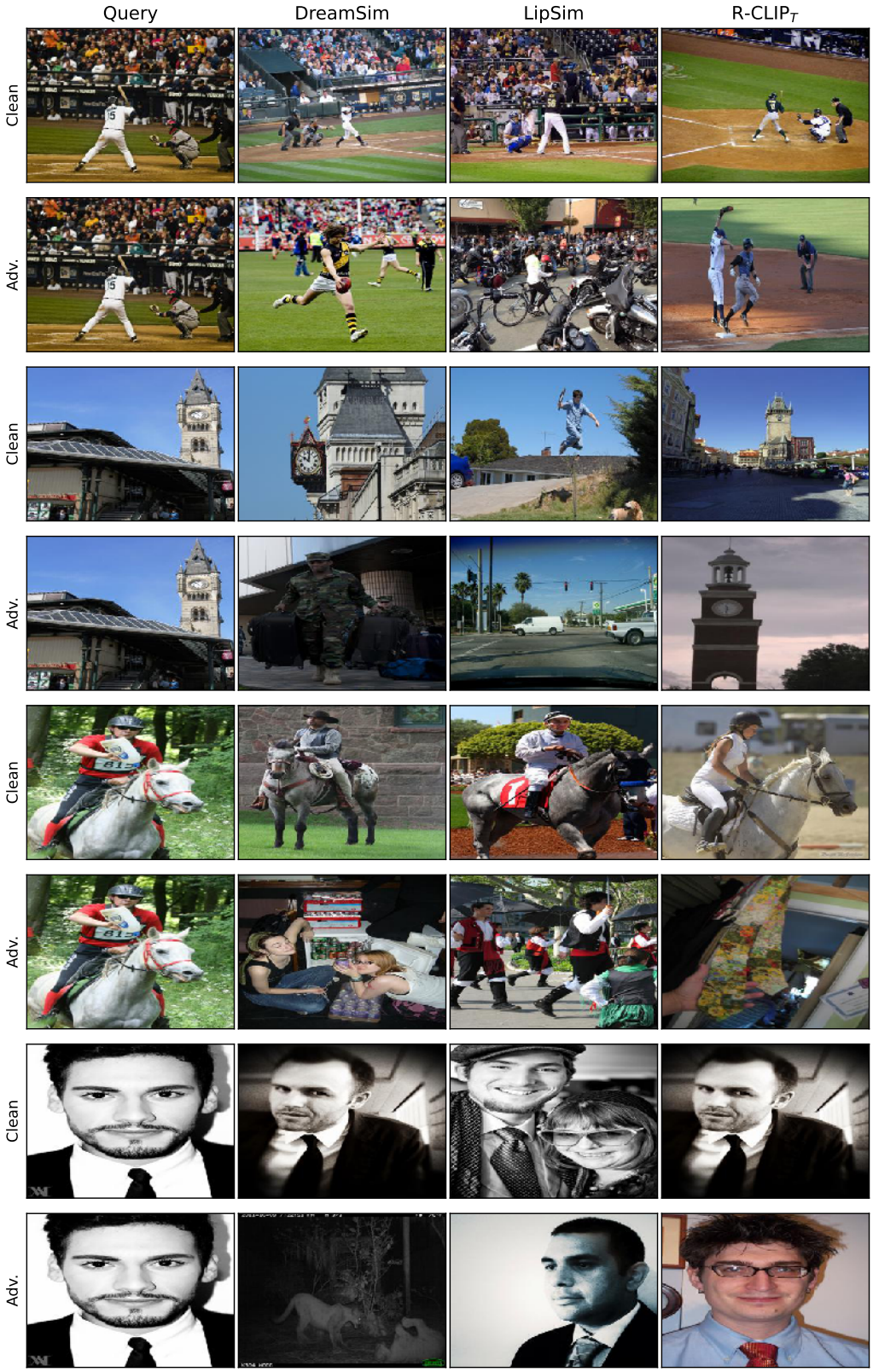

Figure 10. **Clean and adversarial image retrieval on MS-COCO dataset.** Each column shows the nearest neighbour (from 15k random MS-COCO train-set points) to the 'Query' images in the first column. Adversarial images ('Adv.' rows) are generated for ℓ_∞ threat model at $\epsilon = 2/255$ by maximizing the embedding loss of the respective vision encoders.

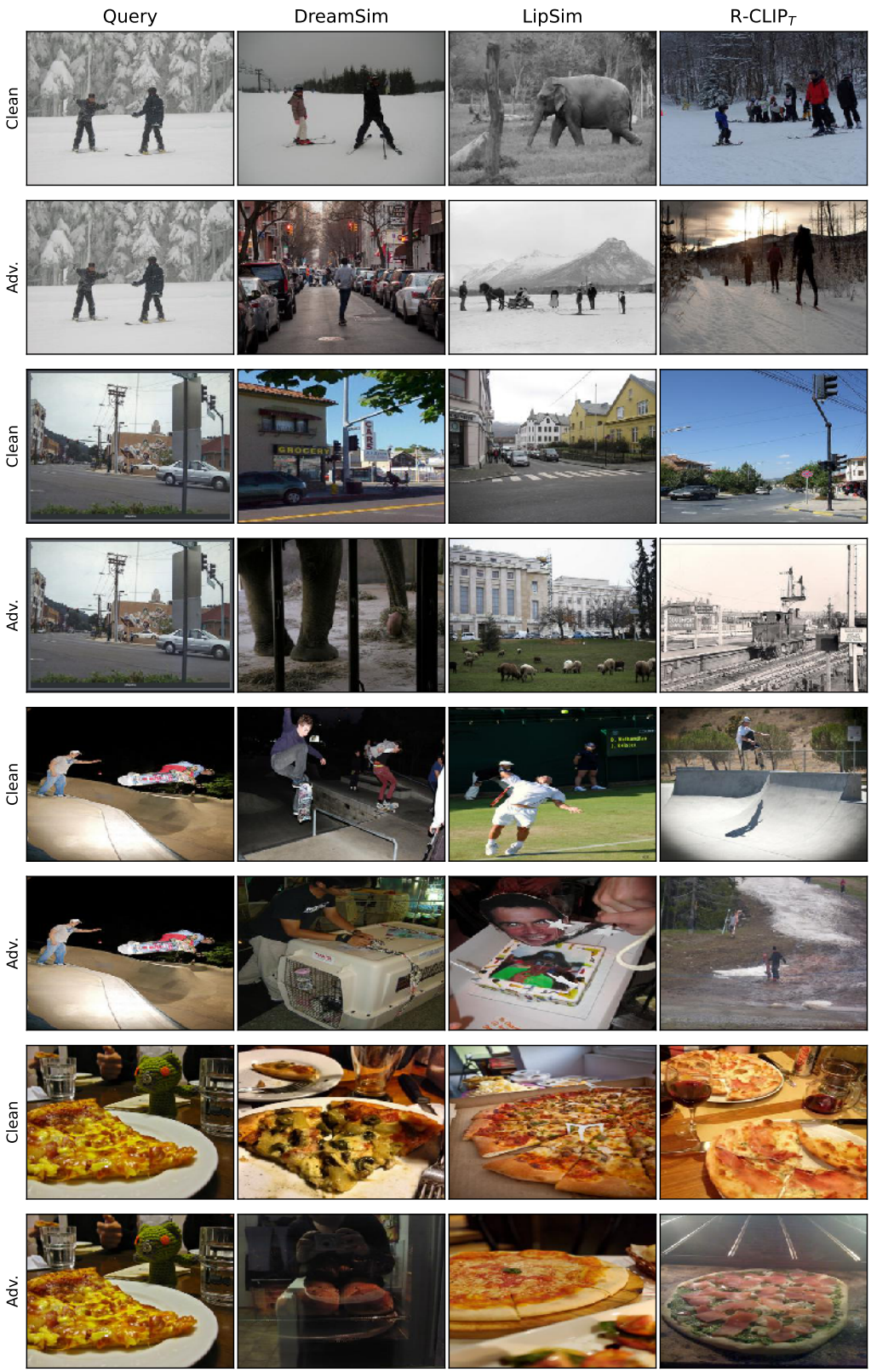

Figure 11. **Clean and adversarial image retrieval on MS-COCO dataset.** The overall setup is same as in Fig. 10.