MIRAGE: RAISING THE COST OF MALICIOUS IMAGE EDITING VIA FALSE MODERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The proliferation of AI-powered image editing systems raises serious concerns about personal images being arbitrarily manipulated at scale with minimal effort and a lower entry barrier. Prior works on *image immunization* require access to the model weights and the image editing prompt, which significantly limits their use, especially against powerful commercial black-box image editors such as GPT-Image, Gemini Flash Image (Nano Banana), and Grok Imagine. To address this, we take a system-level view of the problem and identify a previously unexplored protection surface common to most major commercial systems: pre-generation safety moderation. Rather than directly targeting the image-editing model, we target these moderation classifiers to flag immunized images as policy-violating, triggering an automatic refusal. We operationalize this by adding adversarial perturbations to align our image to policy-violating concepts in the representation space. Under the same level of perturbation, our approach, MIRAGE, significantly improves immunization success rate from less than 15% achieved by the best prior technique to more than 88% on multiple closed-source image editing APIs. Our approach is simple, prompt-agnostic, and effective, offering a practical path towards protecting personal images from unauthorized AI-powered editing.

1 INTRODUCTION

Rapid advances in AI-powered image editing have lowered the barrier to realistic, large-scale photo manipulation. Systems such as OpenAI's GPT-Image, Google's Gemini, and xAI's Grok allow anyone to upload any image and apply sophisticated, semantically meaningful edits through natural language prompts. This includes changing a person's appearance, altering their surroundings, or placing them in entirely fabricated scenarios. While serving legitimate creative purposes, they also enable a new class of consent violations; individuals' personal photographs and images can be maliciously edited without their knowledge or agreement, enabling fraud (DiResta & Goldstein, 2024), harassment (Gentleman, 2026), defamation (Boztas, 2026; Perrigo, 2023), political disinformation (Reuters, 2023), and identity manipulation (Ricker et al., 2024), at a scale previously impossible.

In response to this escalating threat, a series of powerful *image immunization* techniques have been proposed, which add small perturbations to an image in order to cause the image-editing models to fail. However, existing techniques suffer from a **lack of transferability**. Popular methods, including PhotoGuard (Salman et al., 2023), EditShield (Chen et al., 2025), and others (Guo et al., 2024; Ozden et al., 2024), require white-box access to the model weights and also the editing prompt that the perpetrator is going to use; an image immunized for one open-source model and for one prompt is not immune to other models and prompts. Consequently, none of the existing techniques can be used against closed-source black-box commercial APIs, as we show in Figure 3. Currently, individuals sharing images and photos they own are left unprotected against even a casual adversary, who can easily access powerful systems from, say, OpenAI, Google, and xAI.

This leads to a fundamental question in making image immunization practical: *is there an image immunization method that can protect against the frontier AI-powered image editors like GPT-Image, Nano Banana, and Grok Imagine?* Ideally, we want a single immunized image to universally work for these powerful systems and for all image editing prompts, simultaneously. As with prior work, e.g., (Salman et al., 2023), the goal is to raise the cost of malicious image editing for the adversary, such that we raise the barrier of entry.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

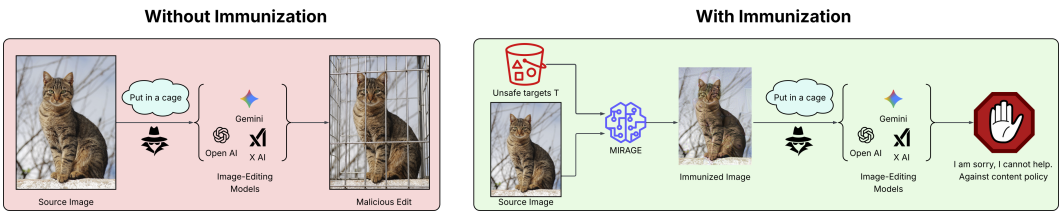


Figure 1: **Overview of the image immunization framework**. *(left)* An adversary can take a source image (without permission) and feed it to a black-box commercial image-editing model with a malicious instruction. The model will generate an image (in which the cat appears in a cage) as instructed. *(right)* To protect against such edits, an image owner could *immunize* their image. Our immunization pipeline takes a set of unsafe target inputs $\mathcal{T}$ (which trigger moderation) along with the source image. The pipeline slightly perturbs the source image so that submitting the immunized image to the editing system yields a refusal and a content-policy violation message..

One might attempt to use derivative-free optimization methods designed for black-box adversarial attacks to directly immunize images against frontier API systems, e.g., (Andriushchenko et al., 2020). However, such methods have large, dimension-dependent query complexity (Duchi et al., 2015). For images, this requires thousands of API calls, which is prohibitively expensive, both in terms of cost (over $50 per image) and time (30 seconds per API call). Alternatively, universal perturbation approaches designed to transfer to any image editors, such as (Ahn et al., 2025; Kim et al., 2025), have been introduced. However, they are ineffective if the ensemble of surrogate models does not include models similar to the target. This is the case for the common commercial, black-box image editing systems as we show in Fig. 3 (qualitative examples are in App. E.1).

To this end, we take a system-level view of current frontier closed-source image-editing systems to identify a fundamental structural property: major commercial APIs refuse service on unsafe images. Before any edits are produced, each system passes the input image through a moderation pipeline; if the image is flagged as policy-violating, the system refuses to generate output entirely, regardless of the prompt. We turn this safety mechanism, originally intended to prevent misuse in specific cases, into a broad defensive tool for image owners. We call this new paradigm *false moderation*: adding perturbations with small visual distortion to an image that cause the safety classifier of a commercial image-editing API to flag the immunized image as policy-violating, producing a refusal that is prompt-agnostic.

The main idea of our method is summarized in Fig. 1. We first investigate the terms of use of major image-editing systems: xAI's Grok (xAI, 2025), Google's Gemini (Google, 2026), OpenAI's GPT-Image (OpenAI, 2025), and extract text descriptions or images associated with policy-violating content categories. We use these to guide the image perturbation to make the image seem unsafe. The resulting perturbation is small in magnitude, and causes the target API to refuse to produce any edit. We term our approach MIRAGE, which stands for **M**oderation **I**nduced **R**esistance **A**gainst **G**enerative **E**diting. In summary, we make the following contributions:

- We identify safety moderation as a novel protection surface against malicious editing using closed-source AI-powered image editing systems. We propose MIRAGE, a method that adds small perturbations causing commercial APIs to refuse to edit an image (Sec. 4).
- On four popular commercial image-editing APIs–OpenAI GPT-Image, Google Gemini, xAI Grok and Ideogram 4.0–we empirically show that MIRAGE achieves immunization success rates of more than 88%, while all prior immunization baselines achieve very low success rates (Sec. 5.2).
- We provide an ablation study demonstrating that each component in MIRAGE is critical for success: (i) using an ensemble of open-source embedding models, (ii) combining both global and local views of the input image, and (iii) the choice of target concept in the terms-of-use (Sec. 5.4).
- We demonstrate our immunization remains largely effective against an adversary who applies standard image processing operations to bypass immunization (Fig. 4). Further, we show that MIRAGE can be seamlessly combined with, for example, PhotoGuard to protect against both local white-box models and proprietary black-box APIs in Appendix D.2.3.

2 PRELIMINARIES

We define two principals: the *adversary*, who seeks to manipulate images posted online without the consent of their respective image owners, and the *image owner*, who wishes to share images without the risk of malicious manipulation.

Threat model: unauthorized manipulation. The adversary has access to an image $x \in \mathcal{X}$, as well as a malicious edit instruction $t \in \mathcal{I}$ and a proprietary black-box AI-powered image editor $f : (\mathcal{X}, \mathcal{I}) \to \mathcal{X} \cup \{\perp\}$, such as GPT-Image(Open AI, 2026), Gemini-3.1-flash-image (Google DeepMind, 2026), or Grok-imagine-image (xAI, 2026). Here $\perp$ denotes a null response. The adversary's goal is to transform x into a target image $y = f(x, t)$ without the consent of the image owner, posing a risk of fraud, harassment, defamation, disinformation, and identity manipulation. In the sequel, we omit the instruction t whenever it is clear from the context. The adversary can optionally pre-process the image, e.g. by applying blurring, cropping, or histogram normalization, or modify the instruction t. In App. D.2, we also consider an adaptive adversary who can spend more computational resources locally to edit the image.

We focus on proprietary APIs since (i) they provide a lower barrier to entry and significantly higher utility to the adversary than local diffusion models and (ii) existing image immunization techniques such as PhotoGuard fail on proprietary image editors as we demonstrate in Figure 3. To the best of our knowledge, MIRAGE is the first to successfully immunize images against black-box, proprietary image editors. Further, MIRAGE can be seamlessly combined with PhotoGuard, for example, for additional protection against white-box models (Appendix D.2.3).

Problem formulation: image immunization. The image owner wants to post online the original image x, knows that there is an adversary attempting to manipulate the image in a harmful way, but does not know which proprietary image editor f the adversary can access and which instruction t the adversary will use. Further, neither the adversary nor the image owner knows the inner workings of each proprietary model other than what is publicly released. The image owner's goal is to **raise the cost of malicious AI-powered image editing** by posting a perturbed version $\hat{x}$ instead of x such that (i) it is visually close to the original image as measured by some image distance metric $d(x, \hat{x})$, and (ii) it renders image editing ineffective, e.g., $d(f(\hat{x}), f(x))$ is large or $f(\hat{x}) = \perp$. The goal in such a scenario is to force the adversary to spend additional resources, making them less incentivized to maliciously edit images without prior consent.

3 RELATED WORK

Our work is at the intersection of three lines of research: (i) image immunization against AI-powered editing, (ii) adversarial attacks repurposed for defensive goals, and (iii) black-box attacks on proprietary models. We describe image immunization here, and defer the other related works to App. A.

As image editing models become more powerful, the potential of their misuse has also increased. As a countermeasure, a recent line of works has sought to disrupt or raise the cost of malicious AI-powered image editing by harnessing adversarial examples (Goodfellow et al., 2015) for diffusion models to disrupt the editing process. PhotoGuard (Salman et al., 2023) was the first method in this line of works, proposing to add a small, optimized perturbation δ to the original image such that the edited output $f(x + \delta)$ is degraded. However, this requires computing $\nabla_\delta f(x + \delta)$, which is not only computationally expensive but also limits the application to white-box image editors. Follow-up works aimed to improve the computational efficiency (Rombach et al., 2022; Choi et al., 2025; Kim et al., 2025; Ahn et al., 2025) as well as efficacy of the perturbation (Chen et al., 2025; Guo et al., 2024; Zhang et al., 2025a; Lo et al., 2024; Dong et al., 2025) through various techniques. We detail these in Appendix A. However, note that all these methods require complete white-box access to the model that the image is to be immunized against. Further, unlike adversarial examples against classification models, we note that these perturbations do not transfer from one diffusion model to another. Recent works have attempted to regularize the perturbation to encourage such transfer (Zhang et al., 2026a), however, as we show in Section 5, such perturbations are ineffective against production black-box image editing systems. This is especially concerning since most powerful image-editing models are black-box, i.e., the user only has API access to these. As a result, existing image immunization methods cannot protect against such models, as we show in Figure 3.

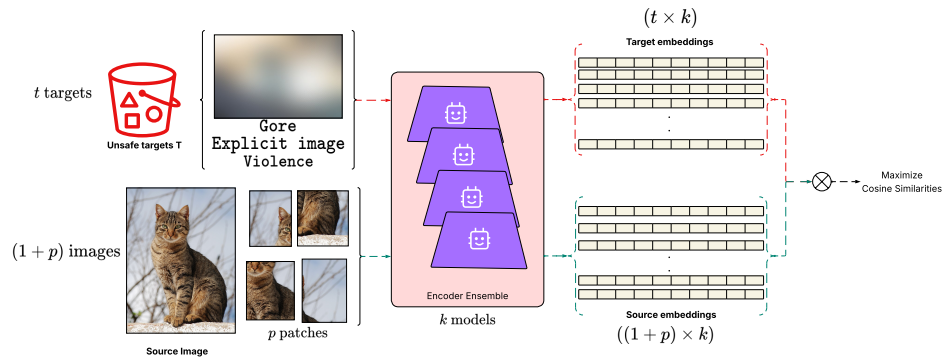


Figure 2: MIRAGE**: Immunization objective** We maximize the cosine similarity between (patch-wise and global) embeddings of our image and target concepts under an ensemble of encoders.

4 METHODOLOGY: MIRAGE

We now introduce our method MIRAGE, which stands for **M**oderation **I**nduced **R**esistance **A**gainst **G**enerative **E**diting. MIRAGE is motivated by the observation that black-box proprietary image editors are equipped with a safety moderator which refuses to produce output when the input image is flagged as policy-violating for a myriad of reasons. We propose to falsely trigger such moderators to achieve image immunization. This is advantageous for two reasons: (i) it is easier to design the perturbation using white-box surrogate models and the resulting perturbations readily generalize to a black-box moderator, and (ii) API providers are incentivized to tolerate moderation false positives to minimize their legal liabilities (Reuters, 2026). We describe our method visually in Fig. 2.

4.1 MOTIVATION

As discussed in Section 3, existing immunization techniques require the gradient $\nabla_x f(\cdot)$ of the image-editing model. Since we focus on black-box image-editing systems where we only access $f(\cdot)$ through an API, we take an alternative approach. Most proprietary systems have a *publicly accessible policy* on acceptable use (xAI, 2025; Google, 2026; OpenAI, 2025). These policies prohibit, e.g., editing or generating copyrighted material, sexually explicit images, and images depicting graphic violence or gore. Such policies are enforced using a content moderation model $\mathcal{C}(\cdot)$ which operates both on the input image x and the output $f(x)$ of the system, and blocks the output if the moderation score is high for either. Using this moderation to our advantage, we propose MIRAGE, which produces perturbed images $\hat{x}$ such that the content moderation system is falsely triggered on the perturbed input agnostic to the text prompt. This can make the MIRAGE-perturbed image immune to unauthorized manipulation by any image-editing system with safety moderation.

An astute reader might notice that we are kicking the can down the road; the moderation model $\mathcal{C}$ is also typically a black-box model, and producing a perturbed image $\hat{x}$ which can maximize $\mathcal{C}(\hat{x})$ is also a challenging task to do efficiently. However, notice that this is akin to finding a transferable adversarial example for a classification task, where the target class corresponds to concepts which are prohibited in the terms-of-use of the black-box APIs. Similar problems have been well-studied in the literature, and prior work has shown that using an ensemble of strong (white-box) classifiers during search can significantly boost the transferability of adversarial examples to black-box classifiers (Chen et al., 2024). There are many such strong, diverse, general-purpose image classifiers available publicly. For example, an image-text encoder like OpenCLIP can be converted into a binary classifier. We note that a concurrent work of Zhang et al. (2026b) has also proposed using an ensemble of OpenCLIP (Cherti et al., 2023) models to construct adversarial examples, but for jailbreaking tasks of misleading black-box VLMs to produce harmful images, e.g., for identity manipulation and NSFW evasion. However, the techniques from Zhang et al. (2026b) (denoted No GL in Figure 5c) cannot directly work in our setting, degrading significantly for high resolution images, achieving 31% immunization rate for Grok, whereas MIRAGE achieves 100%. This difference critically relies on two novel components we added: global and local views in Eq. (2) and reduced memory overhead from dropout and gradient caching in Eq. (3).

4.2 CONSTRUCTING THE IMMUNIZED IMAGE

We construct an immunized image $\hat{x} = x + \delta$ by optimizing a perturbation δ that pushes $\hat{x}$ into the *unsafe region* of the embedding spaces used by content-safety moderators of proprietary image-editing APIs, causing them to refuse the requested edit. We emphasize that we assume no access to the actual content moderators. First, we identify a set of concepts commonly prohibited across major proprietary image-editing APIs by surveying the terms-of-service for xAI, Google and OpenAI. We then scrape images from Wikimedia corresponding to these concepts. We denote the resulting target set of images and texts as $\mathcal{T}$. We detail the selection of these targets in App. B. In Figure 5d, we find that sexually explicit content is more strictly moderated, and hence a better choice as a target. Finally, to construct the perturbation δ, we maximize the similarity of $x + \delta$ with $\mathcal{T}$ in the embedding space of an ensemble of open-source image and text embedding models $\{\mathcal{F}_i\}_{i=1}^n$, including OpenCLIP (Cherti et al., 2023), DinoV2 (Oquab et al., 2024), and open-source moderation models like ShieldGemma (Zeng et al., 2025). We optimize

$$\max_{\|\delta\|_\infty \leq B} \sum_{i=1}^{n} \phi_i(x + \delta, \mathcal{T}), \tag{1}$$

where ϕ_i is the mean cosine similarity between the embeddings of the perturbed image and the target set (or the unsafe score of $(x + \delta)$ in case of moderation models), and B is a constraint. This optimization is done via PGD (Madry et al., 2018). As discussed in Sec. 4.1, Zhang et al. (2026b) also use CLIP-similarity to create transferable adversarial examples. We use a number of practical techniques to better solve the optimization objective in Eq. (1), detailed below and in App. B. We empirically demonstrate the gain of some of the techniques in Figs. 5b–5d.

Handling larger images. We note that the image encoders of most embedding models have been trained on fixed resolution images of small sizes (e.g. 512×512), yet a practical immunization system must handle high-resolution images. One could simply resize the perturbed image $\hat{x}$. However, this leads to spatially coarse gradients, leading to significant performance degradation (see No GL in Fig. 5c). To address this, we decompose the optimization into *global* and *local* views. Let r denote the encoder's native resolution. The global view ($\hat{x}_{\mathrm{G}} = \texttt{Resize}(\hat{x}, r)$) downscales the full image to r, capturing coarse semantic structure. The local view consists of p patches $\{\hat{x}_{\mathrm{L}}^{(j)}\}_{j=1}^p$, each of resolution r, randomly cropped from $\hat{x}$ at its native resolution, preserving fine-grained detail. For each surrogate ϕ_i, recall from Eq. (1) that $\phi_i(\cdot, \mathcal{T})$ denotes the target alignment score. We extend $\phi_i(\cdot, \mathcal{T})$ to incorporate both views:

$$\phi_i(\hat{x}_{\mathrm{G}}, \mathcal{T}) \; + \; \frac{\lambda}{k} \sum_{j \in \text{top-}k} \phi_i\Big(\hat{x}_{\mathrm{L}}^{(j)}, \mathcal{T}\Big) \;, \tag{2}$$

where $\lambda > 0$ is a scalar hyperparameter that balances global semantic alignment against local patch alignment, and top-k selects the $k \leq p$ patches most semantically aligned with $\mathcal{T}$. This empirically leads to better optimization as opposed to using all patch similarities. The full optimization objective sums Eq. (2) across all n surrogate models, analogous to Eq. (1). Incorporating both views leads to significantly more effective immunization than either view alone in Figure 5c, especially against Grok where the immunization rate increases from 17% to 75% at a perturbation budget of 8/255.

Model dropout and secant gradient caching. Optimizing Eq. (1) involves computing $\nabla_\delta \phi_i(x + \delta, \mathcal{T})$ for each surrogate ϕ_i at each PGD step. This, however, becomes memory-intensive as the ensemble and image resolution grow. To reduce the memory and computational cost of this procedure, we propose to use model dropout: at each optimization step, we select a random subset of models to compute the gradient on. To improve optimization stability for the dropped models at a step, we use an approximation of their gradients. Concretely, for each model ϕ_i, we maintain a cache of the past two most recently computed gradients g_i^{t-1}, g_i^{t-2} and the perturbations at which those gradients were computed $\delta_i^{t-1}, \delta_i^{t-2}$. A first order approximation of the gradient would be

$$\tilde{\nabla}_\delta \phi_i(\delta^t, \mathcal{T}) \;=\; g_i^{t-1} \;+\; \nabla_\delta^2 \phi_i(\delta_i^{t-1}, \mathcal{T})\big(\delta^t - \delta_i^{t-1}\big).$$

Rather than computing the full Hessian $\nabla^2 \phi_i$ at δ_i^{t-1}, we use a rank-1 approximation along the secant direction $\delta_i^{t-1} - \delta_i^{t-2}$, yielding:

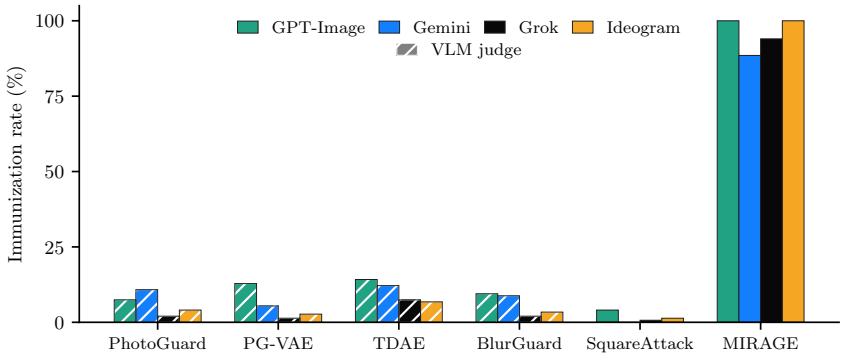


Figure 3: **Immunization rates of baselines and our method across closed-source image editing APIs (OpenAI GPT-Image, Google Gemini, xAI Grok, Ideogram API).** We find that no existing baseline can effectively immunize against proprietary AI-powered image editors. Images immunized using MIRAGE reliably trigger the moderation filters of these systems, leading to the image editing request being refused most of the time. For baseline methods, the success of immunization is judged by Gemini-3.1-Flash with prompts shown in Appendix C.3. The perturbation bound for all methods is $\|\delta\|_\infty \leq \frac{16}{255}$ in these experiments.

$$\tilde{\nabla}_\delta \, \phi_i(\delta^t, \mathcal{T}) \;=\; g_i^{t-1} \;+\; (g_i^{t-1} - g_i^{t-2}) \frac{\langle s_i^t,\; \delta^t - \delta_i^{t-1} \rangle}{\|s_i^t\|_2^2} \;, \tag{3}$$

where $s_i^t = \delta_i^{t-1} - \delta_i^{t-2}$. The second term is a rank-1 curvature correction. It rescales the most recent gradient change by how far the current iterate has moved along the previous secant direction. This can be viewed as a lightweight, one-secant quasi-Newton prediction of the dropped gradient, requiring only two cached gradients and their evaluation points. The approximation is inspired by classical quasi-Newton and secant methods, which estimate curvature from successive gradient differences rather than explicit Hessian computation (Broyden, 1965; Nocedal, 1980; Liu & Nocedal, 1989), and is related to incremental-gradient methods that utilize memories of past component gradients (Schmidt et al., 2017; Defazio et al., 2014; Sohl-Dickstein et al., 2014).

5 RESULTS

For our main experiments, we evaluate MIRAGE against four closed-source image-editing APIs: OpenAI's GPT-Image, Google's Gemini, xAI's Grok, and Ideogram and compare against prior immunization baselines. Existing methods fail entirely in this setting, while MIRAGE achieves high immunization rates across all four APIs (Figure 3). We further analyze the robustness under adversarial pressure (Section 5.3). We also ablate each component of our method and analyze our design choices (Section 5.4). Additional results on adaptive attacks, qualitative results, practical considerations, and detailed ablations are present in App. D and E.

5.1 EXPERIMENTAL SETUP AND BASELINES

Datasets and evaluation. We source test images and prompts from ImgEditBench (Ye et al., 2026), a diverse image editing benchmark, and SHHQ (Fu et al., 2022), a dataset of high-resolution human images with the edit prompt being *"Put this person inside an office"*. Crucially, all images are evaluated at their *full native resolution*—up to 1024x1024—without downsampling prior to immunization or submission to the APIs. Notably, prior work (Salman et al., 2023) usually operates at a resolution of 512x512. Our main results are reported on 100 images drawn from ImgEditBench and 48 images from SHHQ; ablations use the 48-image subset of SHHQ on three APIs. This size is commensurate with prior evaluation (Salman et al., 2023; Zhang et al., 2026a).

We report the *immunization rate*: the fraction of images for which the target API's edit fails. For MIRAGE, this means that the API refuses to respond with an image, hence we measure the number

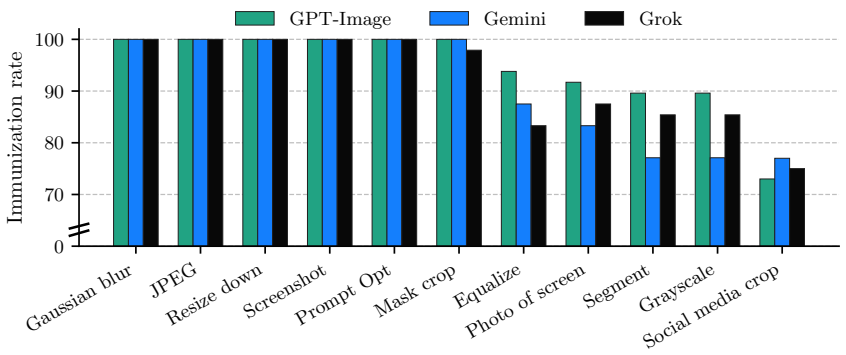


Figure 4: **Robustness of our method under adversarial transforms**. MIRAGE survives most transforms to the image or prompt, achieving immunization rates over 73% in the worst-case.

of requests where a content policy violation message is returned. For the baselines, since APIs might return degraded or semantically unedited outputs rather than explicit refusal, we report VLM-as-a-judge metrics (Ye et al., 2026) to verify that the edit produced was incorrect. We highlight this metric in our plots wherever used. Unless otherwise specified, results are reported at a perturbation budget of $\|\delta\|_\infty \leq 16/255$, commensurate with prior work (Salman et al., 2023; Zhang et al., 2026a).

APIs. We evaluate against four production image-editing APIs: OpenAI GPT-Image-2 (Open AI, 2026), Google Gemini-3.1-Flash-Image (Google DeepMind, 2026), xAI Grok-Imagine-Image (xAI, 2026), and Ideogram 4.0 (AI, 2026). All APIs are accessed via their standard developer endpoints. We note that these systems enforce content moderation both at the input and output stages. While MIRAGE targets the classifier $\mathcal{C}$, which is triggered before any generation is returned to the user, we also report refusals which might arise because of post generation moderation. As with all live systems, moderation sensitivity can change silently over time as providers change their systems. For reporting results, we ran all our experiments on Aug 28, 2026.

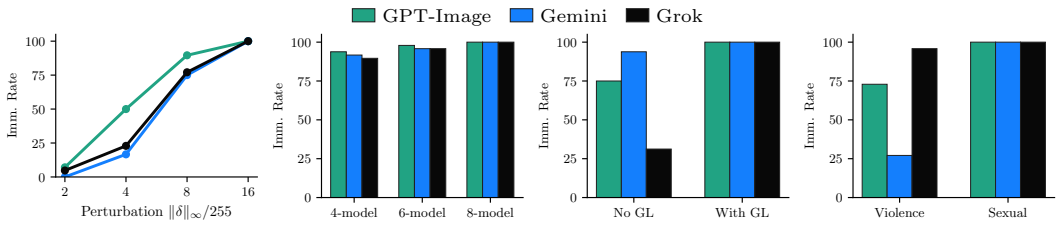
Baselines. We compare against the following immunization methods: white-box attacks PhotoGuard and PhotoGuard-VAE (Salman et al., 2023), BlurGuard (Kim et al., 2025), TDAE (Zhang et al., 2026a), and the black-box SquareAttack (Andriushchenko et al., 2020). Details of these methods and the hyperparameters used can be found in App. C.2.

5.2 MAIN RESULTS

We present our main results in Figure 3. From the plot, it is clear that **all baselines are ineffective in protecting images against unauthorized editing.** Notably, even baselines that explicitly target transferable immunization, such as TDAE (Zhang et al., 2026a) do not protect against image editing by proprietary models. One could contest that the low immunization rates of the baselines are a consequence of our VLM-as-a-judge setup. To address this, we also show qualitative examples of successfully and unsuccessfully immunized images (as determined by the VLM judge) in Figs 14 and 15 in App. E.1. As we see, images are successfully immunized against the local Flux model, leading to degraded output quality. However, on the API models, the output is high quality, albeit not following the instruction in certain cases. This failure mode has also been acknowledged in Zhang et al. (2026a), where weak transfer was observed when perturbations were optimized for weaker models (InstructPix2Pix) and transferred to stronger models (StableDiffusionv3).

The failure of baselines which target the underlying image editing model motivates us to take a systems-level view of this problem. Hence, we turn to attacking the content moderation API. We find that using a black-box adversarial attack (Andriushchenko et al., 2020) to optimize against OpenAI's public moderation API (OpenAI, 2026) can lead to a non-zero refusal rate on their systems, but this does not transfer to other systems as evidenced by near-zero immunization rates.

On the other hand, MIRAGE provides almost perfect protection against GPT-Image and Ideogram 4.0, and immunization rates close to 90% on Gemini and Grok. This difference is because unlike previous baselines, our method does not assume anything about the underlying model powering the



(a) Perturbation budget (b) Ensemble size (c) Global-local views (d) Target category

Figure 5: *(Left)* **Variation of immunization rates with perturbation budget**. Larger perturbation budgets uniformly lead to higher immunization rates. *(Right)* **Ablation results at** $\|\delta\|_\infty = \frac{16}{255}$. (b) Larger ensembles of surrogate image and text encoders improve immunization. (c) Combining global and local views is essential for immunizing high-resolution images. (d) Sexually explicit targets yield higher refusal rates than violent targets. Results with $\|\delta\|_\infty = \frac{8}{255}$ are in App. D.1.

image editor, and works against the content moderation of the system instead. We also show the trade-off between $||\delta||_\infty$ and immunization rate in Figure 5a on SHHQ data. We find that increasing the perturbation budget monotonically leads to higher immunization rates. In Figure 6 (and Figure 13 in the appendix) we show examples of perturbed images – higher perturbation budgets lead to visible artifacts for more effective immunization. However, we note that this degradation similar to baselines— at $||\delta||_\infty = 16/255$, the average LPIPS distance of the immunized image produced by MIRAGE is 0.34, while that of the next-best method, TDAE (Zhang et al., 2026a) is 0.37.

5.3 ROBUSTNESS OF MIRAGE

To test the robustness of our proposed approach, MIRAGE, we measure its immunization performance on the 48-image subset of the SHHQ dataset under adversarial pressure. The adversaries can use off-the-shelf image transforms or prompt optimization aiming to bypass immunization.

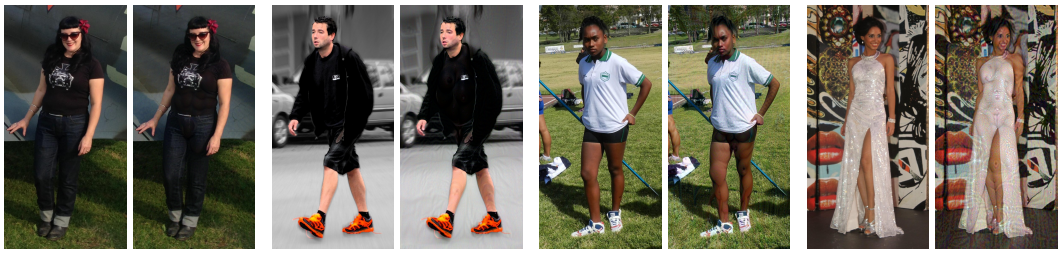
In Figure 4, we present robustness results under image and prompt transforms. The adversarial transforms considered are Gaussian blurring, JPEG compression, resizing the image, cropping the subject to keep the foreground, histogram normalization (equalization), and greyscale. We detail the hyperparameters for these in Appendix D.2. We find that MIRAGE is robust against most such image transformations, with worst case robustness of 73% for GPT-Image with social-media-preview style cropping. In Appendix D.2.2 we also present results under an adaptive attacker who has access to higher computational power, e.g., to run local diffusion models, segmentation models or CLIP-style encoders. We find that MIRAGE provides moderate protection against such adversaries, indicating that it raises the computational cost of malicious image editing. Further, our scheme can be combined with PhotoGuard (App. D.2.3) to simultaneously protect against local diffusion based editors as well as editing APIs with moderation. We show these results in Fig. 12. Finally, we also verify that immunized images can be shared on social media platforms, detailed in App. E.

5.4 ANALYSIS : WHAT IS ESSENTIAL FOR IMMUNIZATION?

We perform an ablation analysis on each component of MIRAGE on the 48-image subset of SHHQ. We use $||\delta||_\infty \leq 16/255$ for these experiments, and report results at $||\delta||_\infty \leq 8/255$ in App. D.1.

Ensemble size. We target an ensemble of models in Eq. (1). For our main experiments, we choose an ensemble of size 8. In Figure 5b we investigate how immunization changes with different ensemble sizes. Across these experiments, we ensure that we cover similar types of models. The list of models can be found in Table 1 in App. C. As shown in Figure 5b, optimizing against a larger surrogate ensemble monotonically leads to better immunization rates. This effect is starker at smaller bounds, where larger ensembles lead to more generalizable perturbations which can mislead the content moderation systems of the proprietary APIs (Fig. 8a in the appendix).

Global-Local (GL) views. In order to deal with high resolution input images, we introduce a mechanism to construct global and local image embeddings according to Eq. (2). For MIRAGE (labeled With GL in the plots), we use 16 random patches to construct the local view, and use top-k



(a) $||\delta||_\infty = 2$ (b) $||\delta||_\infty = 4$ (c) $||\delta||_\infty = 8$ (d) $||\delta||_\infty = 16$

Figure 6: **Qualitative examples of immunized images** We show the original source images and their successfully immunized counterparts at different perturbation budgets (scaled by 255).

as the aggregation function. For the No GL baseline, we simply resize the perturbed image before passing it to each surrogate model $\mathcal{F}_i$ to compute the loss. Using global and local views of the image greatly boosts the immunization rates. This is critical, e.g., for immunizing against Grok, where the immunization rate goes from 31% to 100%, making the method practical.

Targets chosen. For our main experiments, we choose the targets to be a set $\mathcal{T}$ of 5 images and 3 text prompts corresponding to sexually explicit content. In Figure 5d we also experiment with using the same number of images and texts corresponding to graphic violence or gore. As we see, violence is less strictly moderated for GPT-Image and Gemini, leading to lower refusals and immunization rates. However, the content filter for Grok is much more strict for graphic violence or gore, and hence we see a high immunization rate across perturbation bounds.

6 Conclusion

Prior image immunization methods have implicitly assumed that the defender must attack the same generative machinery the adversary uses. This assumption breaks down in the black-box setting where the image-editing models are architecturally diverse, rapidly updated, and entirely inaccessible to gradient-based optimization. All baselines fail against commercial APIs (as we show in Figure 3). Such APIs have a lower barrier of entry and greater efficacy for malicious image editing.

Taking a systems-level view of the problem, we identified a structural property that cuts across most major commercial image-editing systems regardless of their generative architecture: a safety classifier whose refusal is often prompt-agnostic and unconditional. By targeting this shared bottleneck rather than the generative model itself, we sidestep the transferability problem that has plagued prior works. Targeting content-moderation classifiers using standard techniques for transferable adversarial examples is easier for two reasons: (i) several high quality embedding models are open source, and one could use an ensemble of such models to generate transferable examples; (ii) companies are incentivized to tolerate false positives in content moderation to protect against legal liabilities. This gives the defender an asymmetric advantage in this setting. Empirically, MIRAGE achieves close to 90% immunization rate on four commercial APIs (on which every prior immunization baseline achieves very low rates) and remains largely effective under a range of image transformations.

We hope this work shifts the framing of image immunization research: rather than targeting individual generative architectures, responsible use-policies and their enforcement mechanisms represent a broadly shared protection surface for defenders. As AI-powered image manipulation becomes increasingly accessible, practical tools that protect individuals without requiring white-box model access will become correspondingly more important. Further, taking such systems-level views and discovering asymmetric advantages can also help solve other security and privacy problems.

Future Work. While MIRAGE works against commercial image-editing APIs, the robustness against adversaries willing to utilize higher compute (App. D.2) to break the scheme can be improved. Future work can also develop less perceptible perturbations, e.g. by incorporating LPIPS-based constraints into Eq. (1). Finally, more efficient and scalable implementations of MIRAGE can improve the broader accessibility of the method.

AI USE STATEMENT

In this work, we used generative AI tools to help implement experimental code. We did not use generative AI tools to propose or refine hypotheses; design or provide feedback on research methodology or experiments; interpret results; generate synthetic datasets; develop theoretical models or conceptual frameworks; formulate mathematical claims; provide critical ingredients for mathematical proofs; or assist in writing proofs. We additionally used generative AI tools to create or modify scientific figures or images, suggest experimental parameters, create or edit software code and other artifacts, and edit the paper for readability. We reviewed all AI-assisted work, including carefully checking generated code, and take responsibility for the final content, including text, claims, and artifacts produced with AI assistance.

ETHICS STATEMENT

This work introduces MIRAGE, a defensive method for protecting personal images from unauthorized AI-powered editing. We identify the following ethical considerations.

Dual-use risk. Our findings reveal that an ensemble of open-source CLIP-style embedding models transfers effectively to commercial content moderation systems. An adversary could in principle reverse this objective - optimizing perturbations to move away from policy-violating regions in order to evade moderation rather than trigger it. However, evasion and triggering are not symmetric problems: commercial providers deliberately set conservative decision thresholds that tolerate false positives, making evasion substantially harder than triggering. We nonetheless encourage providers to treat perturbation robustness of moderation pipelines as a first-class security concern.

Live system interaction. Our evaluation involved querying four production image-editing APIs (OpenAI GPT-Image, Google Gemini, xAI Grok, Ideogram 4.0) to measure immunization rates. All queries were made through standard developer API endpoints, consistent with each provider's publicly stated terms of service. No attempts were made to extract model weights, circumvent rate limits, or cause disruption beyond the measurement of content policy refusals. Query volume was limited to what was strictly necessary for the reported evaluations.

Datasets and human subjects. Our experiments use two publicly available datasets. SHHQ is released under CC0; it is available for non-commercial research use and we use it solely for evaluation. ImgEditBench is released under the MIT license; we use it as an evaluation benchmark without redistribution. **Protective intent**. The primary goal of this work is defensive: to protect individuals from non-consensual manipulation of their personal photographs by AI-powered editing systems, a harm that most existing immunization methods cannot currently address. We believe this protective intent, combined with the asymmetric difficulty of misusing our findings for evasion, justifies the publication of this work.

REPRODUCIBILITY STATEMENT

We provide all implementation details, including the hyperparameters and algorithmic description of the method in Section 4 and Apps C.1. We will open-source the code upon publication. Note that since the paper works with live systems, some of the exact results might change due to upstream changes to the closed-source APIs.

REFERENCES

Namhyuk Ahn, KiYoon Yoo, Wonhyuk Ahn, Daesik Kim, and Seung-Hun Nam. Nearly zero-cost protection against mimicry by personalized diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28801–28810, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ahn_Nearly_Zero-Cost_Protection_Against_Mimicry_by_Personalized_Diffusion_Models_CVPR_2025_paper.html.

Ideogram AI. Ideogram 4.0. https://ideogram.ai/blog/ideogram-4.0/, 2026.

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: A query-efficient black-box adversarial attack via random search. In *Computer Vision – ECCV 2020*, 2020. URL https://arxiv.org/abs/1912.00049.

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pp. 284–293. PMLR, 2018.

Black Forest Labs. FLUX.2 [klein]: Towards Interactive Visual Intelligence. https://bfl.ai/blog/flux2-klein-towards-interactive-visual-intelligence, January 2026. Blog post. Accessed: 2026-06-08.

Senay Boztas. Dutch far-right party pays damages to court artist after changing image with AI. https://www.theguardian.com/world/2026/jun/13/geert-wilders-pvv-dutch-far-right-party-damages-court-artist-change-image-ai, June 2026.

Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1712.04248.

Charles G. Broyden. A class of methods for solving nonlinear simultaneous equations. *Mathematics of Computation*, 19(92):577–593, 1965.

Huanran Chen, Yichi Zhang, Yinpeng Dong, Xiao Yang, Hang Su, and Jun Zhu. Rethinking model ensemble in transfer-based adversarial attacks, 2024. URL https://arxiv.org/abs/2303.09105.

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pp. 15–26, 2017.

Ruoxi Chen, Haibo Jin, Yixin Liu, Jinyin Chen, Haohan Wang, and Lichao Sun. EditShield: Protecting unauthorized image editing by instruction-guided diffusion models. In *Computer Vision – ECCV 2024*, pp. 126–142. Springer, 2025. doi: 10.1007/978-3-031-73036-8_8.

Minhao Cheng, Thong Le, Pin-Yu Chen, Jinfeng Yi, Huan Zhang, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack:an optimization-based approach, 2018. URL https://arxiv.org/abs/1807.04457.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829. IEEE, June 2023. doi: 10.1109/cvpr52729.2023.00276. URL http://dx.doi.org/10.1109/CVPR52729.2023.00276.

Jianfeng Chi, Ujjwal Karn, Hongyuan Zhan, Eric Smith, Javier Rando, Yiming Zhang, Kate Plawiak, Zacharie Delpierre Coudert, Kartikeya Upasani, and Mahesh Pasupuleti. Llama guard 3 vision: Safeguarding human-ai image understanding conversations. *arXiv preprint arXiv:2411.10414*, 2024.

June Suk Choi, Kyungmin Lee, Jongheon Jeong, Saining Xie, Jinwoo Shin, and Kimin Lee. DiffusionGuard: A robust defense against malicious diffusion-based image editing. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=9OfKxKoYNw.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, 2014.

Renee DiResta and Josh A Goldstein. How spammers and scammers leverage ai-generated images on facebook for audience growth. *arXiv preprint arXiv:2403.12838*, 2024.

Shuai Dong, Jie Zhang, Guoying Zhao, Shiguang Shan, and Xilin Chen. Semantic mismatch and perceptual degradation: A new perspective on image editing immunity. *arXiv preprint arXiv:2512.14320*, 2025. URL https://arxiv.org/abs/2512.14320.

John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks, 2023. URL https://arxiv.org/abs/2309.17425.

Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation, 2022. URL https://arxiv.org/abs/2204.11823.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. DataComp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023. URL https://arxiv.org/abs/2304.14108.

Amelia Gentleman. New claimants seek to sue Elon Musk's xAI after Labour MP's test case. https://www.theguardian.com/technology/2026/jun/05/grok-ai-elon-musk-jess-asato-labour-mp-lawsuit, June 2026. The Guardian. Additional reporting by Jessica Elgot. Accessed June 7, 2026.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6572.

Google. Gemini API Additional Terms of Service. https://ai.google.dev/gemini-api/terms, March 2026. Effective March 23, 2026. Accessed June 7, 2026.

Google DeepMind. Gemini 3.1 Flash Image (Nano Banana 2), 2026. https://deepmind.google/models/gemini-image/flash/.

Zhongliang Guo, Lei Fang, Jingyu Lin, Yifei Qian, Shuai Zhao, Zeyu Wang, Junhao Dong, Cunjian Chen, Ognjen Arandjelović, and Chun Pong Lau. A grey-box attack against latent diffusion model-based image editing by posterior collapse. *arXiv preprint arXiv:2408.10901*, 2024. URL https://arxiv.org/abs/2408.10901.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International conference on machine learning*, pp. 2137–2146. PMLR, 2018.

Xiaojun Jia, Sensen Gao, Simeng Qin, Tianyu Pang, Chao Du, Yihao Huang, Xinfeng Li, Yiming Li, Bo Li, and Yang Liu. Adversarial attacks against closed-source mllms via feature optimal alignment, 2025. URL https://arxiv.org/abs/2505.21494.

Jinsu Kim, Yunhun Nam, Minseon Kim, Sangpil Kim, and Jongheon Jeong. BlurGuard: A simple approach for robustifying image protection against AI-powered editing. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=vritEZz28d.

Dong C. Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, 45:503–528, 1989.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1611.02770.

Ling Lo, Cheng Yu Yeo, Hong-Han Shuai, and Wen-Huang Cheng. Distraction is all you need: Memory-efficient image immunization against diffusion-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24462–24471, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Lo_Distraction_is_All_You_Need_Memory-Efficient_Image_Immunization_against_Diffusion-Based_CVPR_2024_paper.html.

Nicholas A Lord, Romain Mueller, and Luca Bertinetto. Attacking deep networks with surrogate-based adversarial black-box methods is easy. *arXiv preprint arXiv:2203.08725*, 2022.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Anshul Nasery, Edoardo Contente, Alkin Kaz, Pramod Viswanath, and Sewoong Oh. Are robust llm fingerprints adversarially robust? *arXiv e-prints*, pp. arXiv–2509, 2025.

Anshul Nasery, Jonathan Hayase, Creston Brooks, Peiyao Sheng, Himanshu Tyagi, Pramod Viswanath, and Sewoong Oh. Scalable fingerprinting of large language models. *Advances in Neural Information Processing Systems*, 38:125116–125152, 2026.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning (ICML)*, 2022.

Jorge Nocedal. Updating quasi-newton matrices with limited storage. *Mathematics of Computation*, 35(151):773–782, 1980.

Open AI. GPT Image 2, 2026. https://developers.openai.com/api/docs/guides/image-generation.

OpenAI. Usage Policies. https://openai.com/policies/usage-policies/, October 2025. Effective October 29, 2025. Accessed June 7, 2026.

OpenAI. omni-moderation Model. https://developers.openai.com/api/docs/models/omni-moderation-latest, 2026. OpenAI API documentation. Accessed: 2026-06-08.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Tarik Can Ozden, Ozgur Kara, Oguzhan Akcin, Kerem Zaman, Shashank Srivastava, Sandeep P. Chinchali, and James M. Rehg. DiffVax: Optimization-free image immunization against diffusion-based editing. *arXiv preprint arXiv:2411.17957*, 2024. URL https://arxiv.org/abs/2411.17957.

Billy Perrigo. How to Spot an AI-Generated Image Like the 'Balenciaga Pope'. https://time.com/6266606/how-to-spot-deepfake-pope/, March 2023.

Xavier Pleimling, Sifat Muhammad Abdullah, Gunjan Balde, Peng Gao, Mainack Mondal, Murtuza Jadliwala, and Bimal Viswanath. Off-the-shelf image-to-image models are all you need to defeat image protection schemes, 2026. URL https://arxiv.org/abs/2602.22197.

Reuters. Fact Check: Online posts reporting explosion near Pentagon on May 22, 2023 are false. https://www.reuters.com/article/fact-check/online-posts-reporting-explosion-near-pentagon-on-may-22-2023-are-false-idUSL1N37J2QJ/, May 2023.

Reuters. Grok's AI image generation tool violated Canadian privacy law, watchdog says. https://www.reuters.com/business/media-telecom/groks-ai-image-generation-tool-violated-canadian-privacy-law-says-watchdog-2026-06-11/, June 2026.

Jonas Ricker, Dennis Assenmacher, Thorsten Holz, Asja Fischer, and Erwin Quiring. Ai-generated faces in the real world: A large-scale case study of twitter profile images. In *Proceedings of the 27th International Symposium on Research in Attacks, Intrusions and Defenses*, pp. 513–530, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious AI-powered image editing. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29894–29918. PMLR, 2023. URL https://proceedings.mlr.press/v202/salman23a.html.

Rylan Schaeffer, Dan Valentine, Luke Bailey, James Chua, Cristobal Eyzaguirre, Zane Durante, Joe Benton, Brando Miranda, Henry Sleight, Tony Wang, et al. Failures to find transferable image jailbreaks between vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 44669–44704, 2025.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1–2):83–112, 2017. doi: 10.1007/s10107-016-1030-6.

Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. Glaze: Protecting artists from style mimicry by text-to-image models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 2187–2204. USENIX Association, 2023. URL https://www.usenix.org/conference/usenixsecurity23/presentation/shan.

Shawn Shan, Wenxin Ding, Josephine Passananti, Stanley Wu, Haitao Zheng, and Ben Y. Zhao. Nightshade: Prompt-specific poisoning attacks on text-to-image generative models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 807–825, 2024. doi: 10.1109/SP54263.2024.00207. URL https://arxiv.org/abs/2310.13828.

Zedian Shao, Hongbin Liu, Yuepeng Hu, and Neil Zhenqiang Gong. Leave my images alone: Preventing multi-modal large language models from analyzing images via visual prompt injection. *arXiv preprint arXiv:2604.09024*, 2026.

Jascha Sohl-Dickstein, Ben Poole, and Surya Ganguli. Fast large-scale optimization by unifying stochastic gradient and quasi-newton methods. In *International Conference on Machine Learning*, 2014.

xAI. xAI Acceptable Use Policy. https://x.ai/legal/acceptable-use-policy, January 2025. Effective January 2, 2025. Accessed June 7, 2026.

xAI. Grok Imagine API, 2026. https://x.ai/news/grok-imagine-api.

Jiashu Xu, Fei Wang, Mingyu Ma, Pang Wei Koh, Chaowei Xiao, and Muhao Chen. Instructional fingerprinting of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3277–3306, 2024.

Yang Ye, Xianyi He, Zongjian Li, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, Li Yuan, et al. Imgedit: A unified image editing dataset and benchmark. *Advances in Neural Information Processing Systems*, 38, 2026.

Wenjun Zeng, Dana Kurniawan, Ryan Mullins, Yuchi Liu, Tamoghna Saha, Dirichi Ike-Njoku, Jindong Gu, Yiwen Song, Cai Xu, Jingjing Zhou, Aparna Joshi, Shravan Dheep, Mani Malek, Hamid Palangi, Joon Baek, Rick Pereira, and Karthik Narasimhan. Shieldgemma 2: Robust and tractable image content moderation, 2025. URL https://arxiv.org/abs/2504.01081.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, October 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhai_Sigmoid_Loss_for_Language_Image_Pre-Training_ICCV_2023_paper.html.

Jialong Zhang, Zhongshu Gu, Jiyong Jang, Hui Wu, Marc Ph Stoecklin, Heqing Huang, and Ian Molloy. Protecting intellectual property of deep neural networks with watermarking. In *Proceedings of the 2018 on Asia conference on computer and communications security*, pp. 159–172, 2018.

Jie Zhang, Shuai Dong, Shiguang Shan, and Xilin Chen. Dual attention guided defense against malicious edits. *arXiv preprint arXiv:2512.14333*, 2025a. URL https://arxiv.org/abs/2512.14333.

Jie Zhang, Shuai Dong, Shiguang Shan, and Xilin Chen. Towards transferable defense against malicious image edits. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026a. URL https://www.computer.org/csdl/journal/tp/5555/01/11421009/2eApjoNgEve. Early Access / PrePrints.

Jie Zhang, Pura Peetathawatchai, Florian Tramèr, and Avital Shafran. Laundering ai authority with adversarial examples, 2026b. URL https://arxiv.org/abs/2605.04261.

Peng-Fei Zhang, Guangdong Bai, and Zi Huang. Maa: Meticulous adversarial attack against vision-language pre-trained models, 2025b. URL https://arxiv.org/abs/2502.08079.

Vadim Ziyadinov and Maxim Tereshonok. Low-pass image filtering to achieve adversarial robustness. *Sensors*, 23(22):9032, 2023.

A Extended Related Works

A.1 Image Immunization

PhotoGuard (Salman et al., 2023) proposes adding a small perturbation δ to the original image x such that when the perturbed image is passed to the diffusion-based editing model f, the output is visually degraded. To compute the perturbation, PhotoGuard performs PGD (Madry et al., 2018) to minimize $||f(x+\delta) - y_{\text{tgt}}||_2^2$, where y_{tgt} is an arbitrary target image chosen by the image owner (e.g. a completely black image). Crucially, this requires computing $\nabla_\delta f(x+\delta)$, limiting its application to white-box image editors. One downside of PhotoGuard is the cost of backpropagating through the entire diffusion model f. To mitigate the cost, prior works (Salman et al., 2023; Chen et al., 2025; Guo et al., 2024) have focused on the popular latent diffusion based image editing models (Rombach et al., 2022) of the form $f(x) = g(\mathcal{E}(x))$, where $\mathcal{E}(\cdot)$ is an encoder of a VAE and $g(\cdot)$ is the diffusion model in the lower-dimensional latent space. Concretely, an alternate version of PhotoGuard also from (Salman et al., 2023) tries to maximize the distance between $x+\delta$ and x in the latent space of the VAE, i.e., $||\mathcal{E}(x+\delta) - \mathcal{E}(x)||_2$, used for the latent diffusion model of interest. Follow-up works have modified this objective to collapse the posterior of the VAE (Guo et al., 2024). A related line of works aim to create a mismatch between attention matrices (Zhang et al., 2025a; Lo et al., 2024) or intermediate features (Dong et al., 2025) of $x+\delta$ and x in the diffusion transformer used by f.

Another source of wasted resources is that the perturbation δ needs to be reinvented for each image x. More recent works address this by borrowing techniques from the adversarial attacks literature, such as using augmentations (Chen et al., 2025), masking (Choi et al., 2025), universal adversarial attacks to get a single perturbation δ that works across multiple images (Zhang et al., 2026a; Kim et al., 2025), or using a learnt network (Ahn et al., 2025) to generate $\hat{x}$ from x.

A.2 Adversarial attacks for good

The field of image immunization (Salman et al., 2023) uses adversarial attacks to prevent malicious edits to user images. There have been other lines of work reframing techniques from adversarial machine learning for defending against unauthorized AI usages. Nightshade (Shan et al., 2024) and Glaze (Shan et al., 2023) leverage data poisoning attacks to protect style mimicry of artists' work by generative AI models. Model fingerprinting (Zhang et al., 2018; Xu et al., 2024; Nasery et al., 2026; 2025) uses backdoor attacks to detect misuse of open-weights models. Other works (Shao et al., 2026) also construct adversarial examples on white-box visual question answering (VQA) models to prevent them from answering questions on a user's private images. Although image immunization is posed as an adversarial attack problem, our key technique is novel: turning safety moderation of proprietary models against those proprietary systems to make image sharing safe.

A.3 Attacking proprietary models

Popular approaches to attack proprietary machine learning APIs rely on either black-box optimization techniques (Andriushchenko et al., 2020; Cheng et al., 2018) or surrogate-based transfer (Liu et al., 2017; Lord et al., 2022; Zhang et al., 2025b; Jia et al., 2025). However, these approaches face a fundamental scalability challenge when applied to image immunization. Black-box gradient estimation methods such as NES (Ilyas et al., 2018), and ZOO (Chen et al., 2017) typically require thousands to tens of thousands of queries even for targeted attacks in the relatively simpler classification setting (Andriushchenko et al., 2020), which involves estimating the gradient $\nabla_\delta \mathcal{L}$ of a scalar loss signal (a logit or hard label).

Commercial image-editing endpoints such as those built around GPT-Image and Google Gemini incur latencies of 15-30 seconds per forward pass, meaning even modest budget of 1,000 queries translates to roughly 4-8 hours of wall-clock time per image, clearly impractical for end users seeking real-time protection. Prior work on hard-label black-box attacks (Cheng et al., 2018; Brendel et al., 2018) has made progress on query efficiency, but still remains intractable for frontier image generation models; it is evaluated only in classification settings with fast, cheap oracles.

For large vision-language models (VLMs), prior work has further shown that finding transferable adversarial examples is challenging (Schaeffer et al., 2025), prompting more sophisticated ensemble-based approaches with intermediate feature alignment (Zhang et al., 2025b; Jia et al., 2025). The

closest approach to ours is a concurrent work (Zhang et al., 2026b) which uses a CLIP ensemble to construct perturbations which can *bypass the moderation* and decision making pathways of commercial VLMs. This enables the production of output that is policy-violating, e.g. producing fake news about real people. Our work on the other hand perturbs input images which induce content-moderation filters to *refuse* outputs on otherwise "benign" images.

B Implementation Details

We first list the implementation details for MIRAGE

Target concept set. We first identify a set of concepts commonly prohibited across major proprietary image-editing APIs. To do so, we conduct a systematic survey of the terms-of-use of xAI's Grok (xAI, 2025), Google's Gemini (Google, 2026), OpenAI's GPT-Image (OpenAI, 2025), and extract the concepts that are universally prohibited, including sexually explicit imagery, graphic violence, and copyrighted material. For each concept, we collect a set of publicly available images and text captions that instantiate it. We scraped the Bad image list from WikiMedia to obtain 800 target images. For each moderation category the CLIP image embeddings are tightly clustered. Ablation studies choosing small numbers of target images randomly found that the variance in the mean embedding was small. We hence picked 5 target images randomly per moderation category. We denote the resulting target set of images and texts as $\mathcal{T}$. In Figure 5d, we find that sexually explicit content is more strictly moderated, and hence a better choice as a target.

Models used To maximize the chance of transfer to an unknown content moderation model $\mathcal{C}$, we use a diverse set of surrogate models $\mathcal{F}_i$ in Eq. (1). We use 8 transformer-based CLIP-style models trained on diverse data (Cherti et al., 2023; Zhai et al., 2023; Fang et al., 2023; Gadre et al., 2023). We also use self-supervised image encoders (Oquab et al., 2024) to produce more robust perturbations. From both these types of models, we obtain the embeddings of the perturbed image, as well as the embeddings of the target images and/or texts, and maximize the cosine similarity between the two. Finally, we also leverage open-source moderation models (Zeng et al., 2025; Chi et al., 2024) which return a scalar between $[0, 1]$ denoting how unsafe an image is. We aim to maximize this score for our perturbed image. Figure 5b shows how more models in the ensemble helps immunization, but also the gain diminishes as we increase the ensemble size and at larger perturbation bound.

Augmentations and optimization Following prior work producing robust adversarial examples (Athalye et al., 2018; Zhang et al., 2026b), we apply stochastic augmentations (augmentation α drawn uniformly at random from a set $\mathcal{A}$, i.e., $\alpha \sim \mathcal{A}$) to $x + \delta$ before passing it to ϕ_i, replacing $\phi_i(x + \delta, \mathcal{T})$ with $\phi_i(\alpha(x + \delta), \mathcal{T})$ in Eq. (1). The augmentation set $\mathcal{A}$ includes random crops, resizing, gaussian blur, JPEG compression, and flips. We sample one augmentation per step, and use a straight-through gradient estimator for non-differentiable augmentations (eg. JPEG compression). We optimize Eq. (1) using PGD for 5,000 steps with iterated FGSM (Goodfellow et al., 2015).

C Experimental Details

C.1 Hyperparameters

Table 1 lists the full set of hyperparameters used for MIRAGE in all main experiments. We use PGD with a cosine step-size schedule for 5,000 steps. The surrogate ensemble consists of 8 models spanning CLIP-style vision-language encoders, a self-supervised image encoder (DINOv2), and an open-source moderation model (ShieldGemma-2). Global-local optimization uses 16 randomly sampled patches of resolution 256×256 at the image's native resolution, with top-k mean aggregation ($k = 8$) over patch similarities. Model dropout is applied with probability 0.3, with a minimum of 3 surrogates selected per step.

It takes 10 minutes to immunize a single 1024x1024 image on 1 A40 GPU if we use 5000 steps.

Parameter	Value
Attack	ℓ_∞ PGD
Steps	5000
Step size	1.0
Schedule	Cosine
Objective category	Sexual
EOT samples	1
Patch size	256
Number of local patches	16
Patch sampling mode	Random
Local view weight	0.25
Local pooling	Top-k mean
Local pooling k	8
Surrogate drop probability	0.3
Minimum selected surrogates	3
Surrogate ensemble	`hf_dinov2:facebook/dinov2-base`; `hf_siglip:google/siglip-base-patch16-224`; `open_clip:ViT-B-32/laion2b_s34b_b79k`; `open_clip:ViT-B-16/laion2b_s34b_b88k`; `open_clip:ViT-L-14/datacomp_xl_s13b_b90k`; `open_clip:ViT-L-14/dfn2b_s39b`; `open_clip:ViT-H-14/dfn5b`; `shieldgemma:google/shieldgemma-2-4b-it:sexual`

Table 1: Hyperparameters for MIRAGE used in the main experiments.

C.2 BASELINES

We have the following baselines —

- **PhotoGuard** (Salman et al., 2023): both the diffusion variant, which minimizes $\|f(x+\delta) - y_{\text{tgt}}\|_2^2$, and the VAE variant, PG-VAE, which maximizes $\|\mathcal{E}(x+\delta) - \mathcal{E}(x)\|_2$ in the encoder's latent space.
- **BlurGuard** (Kim et al., 2025): aims to find robust and transferable perturbations by regularizing the spectrum of the noise.
- **TDAE** (Zhang et al., 2026a): augments the PhotoGuard objective with a Hessian-based curvature regularizer designed to encourage transfer across diffusion model architectures.
- **SquareAttack** (Andriushchenko et al., 2020): a query-efficient black-box adversarial attack applied directly to OpenAI's public content moderation API $\tilde{\mathcal{C}}$ (OpenAI, 2026), representing the strongest possible direct black-box attack on an available classifier.

Table 2 lists the hyperparameters used for all baseline methods. For all white-box baselines, we use Flux-Klein 4B (Black Forest Labs, 2026) as the surrogate target model. All methods are run at $\|\delta\|_\infty \leq 16/255$ unless otherwise noted.

Method	Hyperparameters
SquareAttack	Score loss, stripe initialization, random start, $\epsilon = 16/255$, 500 steps, sexual category, threshold 0.5, no early stopping, $p_{\text{init}} = 0.01$, no max-side resizing
BlurGuard	$\epsilon = 16/255$, 200 steps, step size 1.0, learning rate 0.06, $\epsilon_\sigma = 0.0$, σ weighting 10.0, 50 warmup steps, seed 0
PhotoGuard-VAE	$\epsilon = 16/255$, 200 iterations, step size 1.0, 1 gradient repetition, ℓ_∞ norm, constant schedule, minimum step-size factor 0.0
PhotoGuard-Diffusion	$\epsilon = 16/255$, 200 iterations, step size 1.0, 5 gradient repetitions, zero image target, ℓ_∞ norm, constant schedule, minimum step-size factor 0.0
TDAE	$\epsilon = 16/255$, 200 iterations, step size 1.0, zero image target, Hessian step size 2.0, curvature regularizer 1.0, ℓ_∞ norm, constant schedule, minimum step-size factor 0.0

Table 2: Hyperparameters for baseline methods.

C.3 JUDGE PROMPTS

We use Gemini-Flash-3.1 as our judge for edits. We use the following judging prompts —

SHHQ

"You are judging whether two images depict the same person.

Compare the person's visible identity cues, including face shape, facial features, hairline, hair style, skin tone, and any stable identifying details. Ignore pose, crop, lighting, background, image quality, clothing changes, and style changes when they are not identity cues.

Return only a JSON object with: - same-person: true if Image B appears to depict the same person as Image A, false if it appears to depict a different person, null if impossible to determine. - confidence: a number from 0 to 1. - rationale: one concise sentence."

IMGEDIT

"You are judging whether an image edit succeeded.

You will see Image A before editing, Image B after editing, and the edit instruction. Evaluate whether: 1. The edit instruction was followed very faithfully. 2. No extraneous edit was added. 3. The original subject and content that should not change were preserved. 4. Visual quality was preserved.

Return only a JSON object with: - instruction-followed: true or false. - no-extraneous-edits: true or false. - subject-preserved: true or false. - quality-preserved: true or false. - overall-success: true only if all four criteria are true, otherwise false. - score: an integer from 1 to 5, where 5 is near-perfect, 4 has only minor issues, and 3 or lower has meaningful failures. - rationale: one concise sentence of no more than 25 words."

D ADDITIONAL RESULTS

We provide tabular versions of all figures in the main paper for reference and reproducibility.

Table 3 reports the main immunization rates for all baselines and MIRAGE across the four target APIs. Baseline entries other than SquareAttack and MIRAGE are VLM-as-a-judge results, since white-box methods do not produce explicit API refusals but may degrade edit quality. Table 6 reports immunization rates as a function of perturbation budget $\|\delta\|_\infty$ for MIRAGE. Table 5 reports immunization rates under all weak, compound, and strong adversarial transformations; for strong attacks, parenthesized values are VLM-as-a-judge rates where available. Table 4 reports the full ablation results on the 48-image SHHQ subset at both perturbation budgets.

Method	GPT-Image	Gemini	Grok	Ideogram
PhotoGuard	7.4%	10.8%	2.0%	4.1%
PhotoGuard-VAE	12.8%	5.4%	1.4%	2.7%
TDAE	14.2%	12.2%	7.4%	6.8%
BlurGuard	9.5%	8.8%	2.0%	3.4%
SquareAttack	4.1%	0.0%	0.7%	1.4%
Ours	100.0%	88.5%	94.0%	100.0%

Table 3: Main immunization rates for baselines and MIRAGE. Baseline entries other than SquareAttack and MIRAGE are VLM-as-a-judge results.

Ablation	$\|\|\delta\|\|_\infty = 8/255$			$\|\|\delta\|\|_\infty = 16/255$		
	OAI	Gemini	Grok	OAI	Gemini	Grok
No Global-local	66.7%	62.5%	16.7%	75.0%	93.8%	31.2%
6-model	89.6%	62.5%	77.1%	97.9%	95.8%	95.8%
4-model	77.1%	58.3%	52.1%	93.8%	91.7%	89.6%
Violence target	33.3%	2.1%	62.5%	72.9%	27.1%	95.8%
No $\vec{C}$	89.6%	75.0%	77.1%	100.0%	100.0%	100.0%
Full	89.6%	68.8%	75.0%	100.0%	100.0%	100.0%

Table 4: Immunization rates for SHHQ ablations.

Attack	Type	GPT-Image	Gemini	Grok
Mask crop	Weak	100.0%	100.0%	97.9%
Equalize	Weak	93.8%	87.5%	83.3%
Gaussian blur	Weak	100.0%	100.0%	100.0%
Grayscale	Weak	89.6%	77.1%	85.4%
JPEG	Weak	100.0%	100.0%	100.0%
Resize down	Weak	100.0%	100.0%	100.0%
Screenshot	Compound	100.0%	100.0%	100.0%
Photo of screen	Compound	91.7%	83.3%	87.5%
Social media crop	Compound	73.0%	77.0%	75.0%
Foreground white	Strong	89.6% (89.6%)	77.1% (77.1%)	85.4% (85.4%)
Diffusion denoising	Strong	56.2% (70.8%)	64.6% (64.6%)	78.0% (78.0%)
CLIP attack	Strong	8.3% (47.9%)	10.5% (14.6%)	6.2% (12.5%)

Table 5: Immunization rates under adversarial transformations. For strong attacks, parenthesized values are VLM-as-a-judge immunization rates where available.

$\|\delta\|_\infty$	GPT-Image	Gemini	Grok
$2/255$	7.1%	0.0%	4.8%
$4/255$	50.0%	16.7%	22.9%
$8/255$	89.6%	75.0%	77.1%
$16/255$	100.0%	100.0%	100.0%

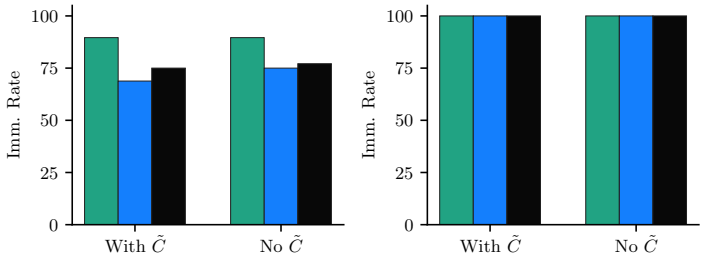
Table 6: Perturbation budget versus immunization rate for MIRAGE.

D.1 Additional Ablations

D.1.1 Using $\tilde{\mathcal{C}}$

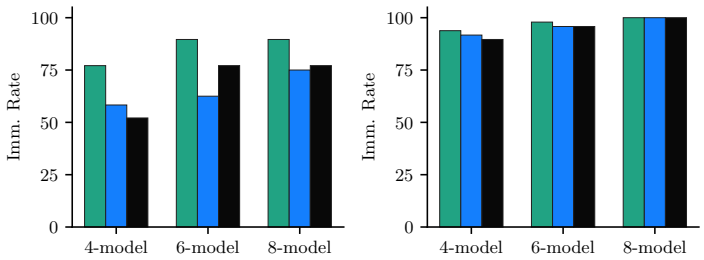
Validation with a public moderation API For early stopping the optimization, we checkpoint $\hat{x}$ every 250 steps and could return the checkpoint with the highest score under a public black-box moderation API, $\tilde{\mathcal{C}}$, from OpenAI (OpenAI, 2026) as the final immunized image. We note two important caveats. Firstly, $\tilde{\mathcal{C}}$ is distinct from the internal moderation model $\mathcal{C}$ used by commercial image-editing APIs even for OpenAI's GPT-Image: an image with high $\tilde{\mathcal{C}}(\hat{x})$ can have a low $\mathcal{C}(\hat{x})$ and still be edited, meaning checkpoint selection via $\tilde{\mathcal{C}}$ does not guarantee transfer.

Secondly, one might ask whether directly optimizing $\hat{x}$ to maximize $\tilde{\mathcal{C}}(\hat{x})$ via black-box optimization is a stronger alternative; as we show in Figure 3, state-of-the-art black-box attacks of SquareAttack (Andriushchenko et al., 2020) against $\tilde{\mathcal{C}}$ produce perturbations that do not reliably transfer to



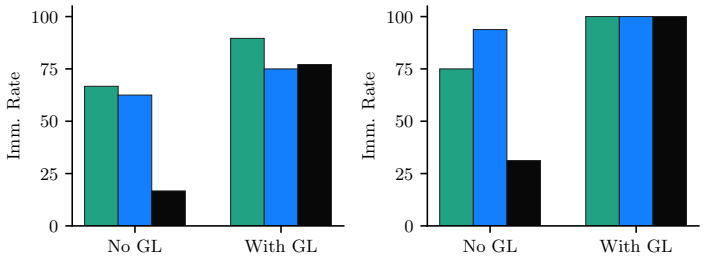
(a) $||\delta||_\infty = 8/255$ (b) $||\delta||_\infty = 16/255$

Figure 7: **Effect of using the surrogate moderation objective for validation**. We find that using a public $\tilde{\mathcal{C}}$ from OpenAI to early stop the optimization can lead to worse immunization on non-OpenAI models (blue and black).



(a) $||\delta||_\infty = 8/255$ (b) $||\delta||_\infty = 16/255$

Figure 8: **Effect of ensemble size on immunization rate.** More models $n \in \{4, 6, 8\}$ in the ensemble objective Eq. (1) makes the immunization generalize better, resulting in improved success rates, but the gain diminishes with larger ensemble and higher perturbation bound.



(a) $||\delta||_\infty = 8/255$ (b) $||\delta||_\infty = 16/255$

Figure 9: **Effect of global-local embeddings on immunization**. Using both global and local views in the objective in Eq. (2) is crucial for immunizing high resolution images.

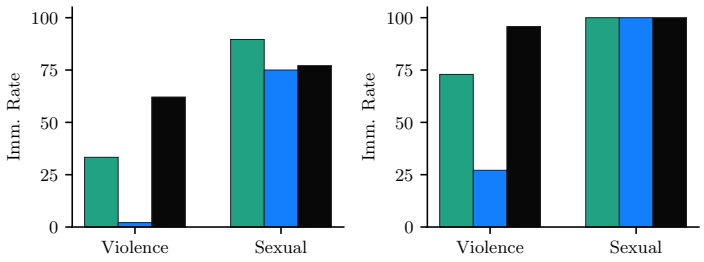
the internal $\mathcal{C}$ of production APIs. In practice, we find that omitting $\tilde{\mathcal{C}}$-based checkpoint selection entirely performs comparably or better (see Figure 7), and we ablate this component in Section 5.

Since a public content moderation API is available from OpenAI (OpenAI, 2026), one could use it to early stop the immunization procedure. However, this leads to poorer generalization across models, as seen in Figure 7.

D.2 Adversarial Evaluation

D.2.1 Weak adversary

1. `Gaussian Blur`: This smooths the image to potentially disrupt high frequency perturbations. This is motivated by successful countermeasure to adversarial examples that tend to add high-frequency perturbations in (Ziyadinov & Tereshonok, 2023). We use a radius of 1 for the Gaussian Blur. We find that MIRAGE survives this transformation perfectly.
2. `JPEG compression`: The adversary compresses the immunized image using JPEG, reducing quality to 50%. This can induce artifacts into the image, however, our immunization scheme is unaffected by this.
3. `Resize down`: The adversary resizes the image to 50% of its original size, potentially erasing fine-grained perturbations. However, MIRAGE sees no degradation in immunization rates under this.
4. `Screenshot`: We simulate an adversary taking a screenshot of the immunized image by resizing the long side to 1080, applying a slight Gaussian blur, then JPEG compression with quality 75. This also does not lead to a degradation in immunization rates.
5. `Prompt Opt`: We empirically tested MIRAGE against an adversary who uses GEPA-style prompt evolution, with 40 prompts per image, to bypass moderation (querying the editing API directly and/or a moderation proxy).



(a) $||\delta||_\infty = 8/255$ (b) $||\delta||_\infty = 16/255$

Figure 10: **Effect of target category on immunization rate**. Using sexually explicit content as the target concept leads to higher refusals as compared to violent imagery.

6. `Subject Crop`: An adversary could crop the image to only keep the foreground subject and eliminate any perturbations contained fully in the background. We obtain the fore-ground bounding boxes from the ground truth SHHQ annotations. We find a small drop in immunization rates for Grok here.
7. `Equalize`: The adversary can perform histogram normalization to remove immunization artifacts. We find that this can reduce the effectiveness of MIRAGE, but the overall rate remains high (83% for Grok).
8. `Photo of a screen`: An adversary can take a photo of a screen displaying the immunized image, in the hopes that this strips away the immunization. We simulate this by noising, blurring, rotating, and increasing the brightness of the image. This degrades the immunization mildly, though the overall rate remains higher than 80%.
9. `Greyscale`: The adversary could convert the image to greyscale to strip away perturbations that depend on colors. We find that this reduces the immunization rate of MIRAGE mildly, with performance on Gemini dropping to close to 77%.
10. `Social Media Crop`: Social media platforms often crop the preview image to be square. We simulate this by cropping a square around the main subject of the image, followed by adjusting the contrast and sharpening the image. This leads to a moderate drop in the immunization rates of MIRAGE, dropping to about 73% for GPT-Image.

D.2.2 STRONG ADVERSARY

In Figure 11, we present results under a strong adversary. We consider four adaptive attacks under this threat model.

1. `Segment`: An adversary could isolate the foreground in a more fine-grained manner as compared to the `Mask Crop` attack in Table 5. They can segment the image with an off-the-shelf model, mask the foreground precisely and paste it onto a white background. We find that this is moderately successful against MIRAGE, leading to immunization rates under 80% for Gemini.
2. `Denoise`: This adversary first uses a local diffusion-based image editing model (Flux-v.2-Klein-4B) to denoise or "purify"(Nie et al., 2022) the image. Instead of using the classical DiffPure approach, we adapt the attack from Pleimling et al. (2026), where the adversary first prompts the local model to denoise the image, and then passes the "denoised" image to the black-box API. This is done in a two stage manner—denoising followed by a call to a proprietary model—since the local model might not be as performant as the proprietary model, and could give degraded outputs when tasked with the actual malicious image edit. We find that this is moderately successful in erasing the immunization, with about 56% refusal rate on GPT-Image after purification. However, taking into account the quality of the image edited by the proprietary systems, we find that the worst immunization rate under attack is 65% for Gemini.
3. `CLIP`: An adversary could try to undo the immunization by minimizing the objective in Eq. (1) using a similar set of public embedding models. We simulate such an adversary by choosing a subset of 4 embedding models, and minimizing the similarity to the target concepts. We run PGD for 2000 steps, with a perturbation bound $||\delta||_\infty \leq 16/255$. This is lower than the resources spent

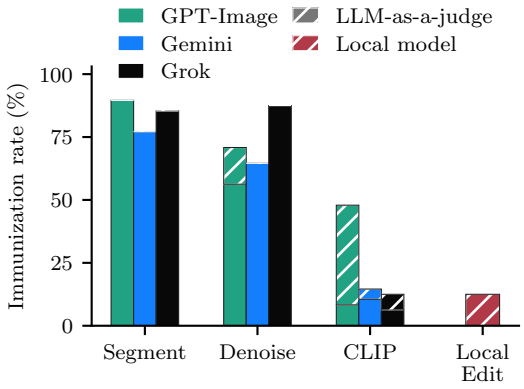


Figure 11: **Immunization rate of our method under stronger adversarial perturbations**. If the adversary can run embedding, generation or segmentation models, they can construct powerful perturbations to bypass the immunization afforded by MIRAGE. VLM-as-a-judge shows the proportion of images that pass the moderation filter but result in significantly distorted output. Local model shows immunization rate against a local diffusion model.

by the defender. However, we find that this attack is pretty successful, reducing the immunization rate to 12.5% for Grok and 14.6% for Gemini. On the other hand, This leads to some degradation in the image, and the VLM-as-a-judge says that the two people are different.

4. `Local Edit`: This is the most potent attack that an adversary can launch, since it completely bypasses the moderation pathway that MIRAGE exploits. Indeed, as we see in the plot this attack can lead to very low immunization rates. However, we note that if the image owner knew what local diffusion model was used by the adversary, they could add a PhotoGuard style objective to Eq. (1) and provide some protection. We describe preliminary results of such compound immunization in App. D.2.3, and find this to be a promising direction.

D.2.3 COMPOUND IMMUNIZATION WITH PHOTOGUARD

As we saw in Section 5.3, a strong adversary with access to a local diffusion model could easily bypass MIRAGE and edit the image locally. If the immunizer knew what local model might be used to edit the images, they could add a PhotoGuard-style objective with the desired local model as one of the surrogates during the immunization procedure. We implement this, and present qualitative results in Fig. 12. In each row, we show the original image, the PhotoGuard perturbation, local and API edits on the PhotoGuard protected image. PhotoGuard can protect against editing by the local model, but not against API edits. We then show the perturbed image found by MIRAGE + PhotoGuard, where a Flux-4B model is used as one of the surrogate models in optimizing Eq. (1). We find that the perturbed image is refused by the image editing APIs. It also leads to degraded image edits by the local model, getting the best of both worlds.

E ADDITIONAL QUALITATIVE RESULTS

Figure 13 shows additional examples of source images and their immunized counterparts at various perturbation budgets.

Uploading Images to Social Media We additionally test whether images immunized by MIRAGE can be uploaded to social media platforms. On the 48-image SHHQ subset, immunized images at $\|\delta\|_\infty = 8/255$ upload without any sensitivity label on X, Instagram, and Reddit (48/48 on each platform). At $\|\delta\|_\infty = 16/255$, all images remain postable on all three platforms and none were rejected outright. We find that X and Reddit mark the resulting post as sensitive media in approximately 25% cases; Instagram applies no such label at either budget.

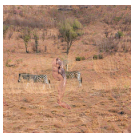
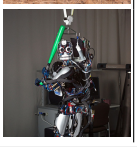
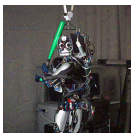
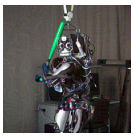
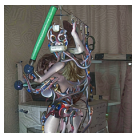
Prompt	Original	PhotoGuard			+MIRAGE		
		Perturbed	Locally edited	API edited	Perturbed	Locally edited	API edited
Edit some mountains in the background.							Refused
Replace the baseball bat with a laser sword.							Refused

Figure 12: Qualitative comparison between PhotoGuard (PG) and MIRAGE on the same edit prompts demonstrates that combining PG and MIRAGE in the perturbation objective provides protection against both open-weight models and black-box APIs, simultaneously and with a single perturbation.



(a) $||\delta||_\infty = 8/255$ (b) $||\delta||_\infty = 8/255$ (c) $||\delta||_\infty = 16/255$ (d) $||\delta||_\infty = 16/255$

Figure 13: Additional qualitative examples of source images and their immunized counterparts.

E.1 QUALITATIVE RESULTS ON BASELINES

From our main results, we see that even baselines that explicitly target transferable immunization, such as TDAE (Zhang et al., 2026a) do not protect against image editing by proprietary models. One could contest that the low immunization rates of the baselines are a consequence of our VLM-as-a-judge setup. To address this, we also show qualitative examples of successfully and unsuccessfully immunized images (as determined by the VLM judge) in Figs. 14 and 15. In each of the figures, we present three successfully immunized images (top three rows) and three immunization failures. As we see, images are successfully immunized against the local Flux model, leading to degraded output quality. However, on the API models, the output is high quality, albeit not following the instruction in certain cases.

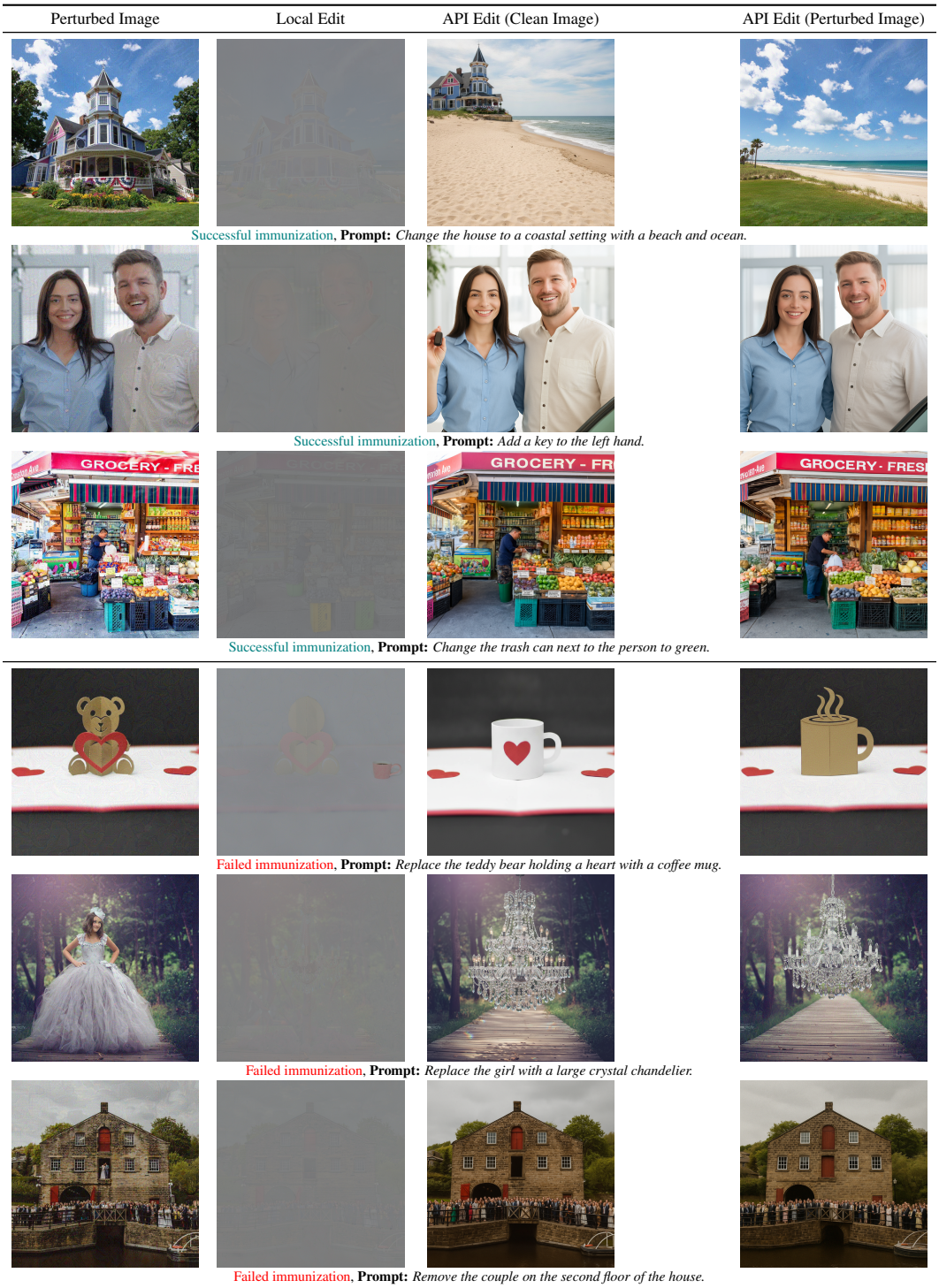


Figure 14: **Qualitative PhotoGuard (Salman et al., 2023) examples**. Each row shows the perturbed image, the local white-box edit, the API edit of the clean image, and the API edit of the perturbed image. The top three rows are successfully immunized, where the VLM judge determines that the perturbed API edit does not follow the instruction, while the bottom three rows are immunization failures. We see that the VLM-as-a-judge is consistent with human judgment for the immunized results on the rightmost column.

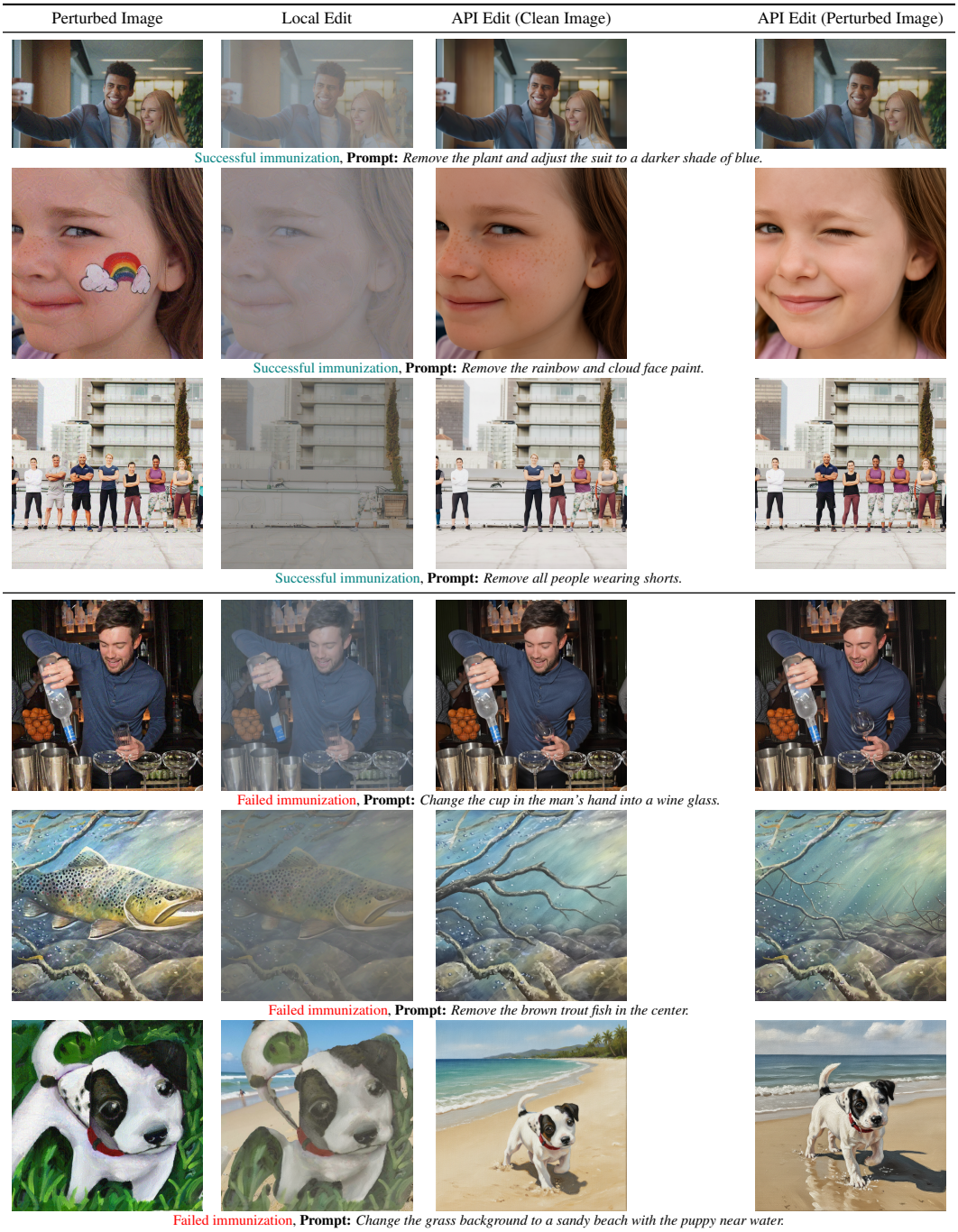


Figure 15: **Qualitative TDAE (Zhang et al., 2026a) examples** Each row shows the perturbed image, the local white-box edit, the API edit of the clean image, and the API edit of the perturbed image. The top three rows are successfully immunized, where the VLM judge determines that the perturbed API edit does not follow the instruction, while the bottom three rows are immunization failures. We see that the VLM-as-a-judge is consistent with human judgment for the immunized results on the rightmost column.