IMMUNE ESCAPE IN MODEL IMMUNIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Efficient fine-tuning techniques have substantially reduced the cost of adapting open-weight models, enabling malicious actors to readily fine-tune them toward harmful capabilities. In response, model immunization modifies a pretrained checkpoint to resist subsequent adaptation toward designated harmful capabilities while preserving its utility and adaptability for benign tasks. However, we observe that existing immunization methods substantially overfit to the specific configurations used during immunization, particularly the images, textual prompts, and adaptation strategy the surrogate attacker is assumed to use. This fragility is obscured by existing evaluations, which largely reuse the same datasets and attack configurations employed during immunization. To rigorously evaluate model immunization, we develop ESCAPE, an attack framework that generates diverse variants of harmful adaptation designed to circumvent immunization. Specifically, ESCAPE misaligns the attacker from the configuration simulated during immunization along three axes: harmful data, model initialization, and adaptation strategy. We evaluate ESCAPE against three existing model immunization methods across diverse tasks and modalities, spanning text-to-image generation, image classification, and language models. Our results show that ESCAPE substantially weakens immunization across all three modalities and, in many cases, largely erases the protection that immunization provides over the un-immunized model.

1 INTRODUCTION

Open-weight release is now standard practice (Kapoor et al., 2024), and a released checkpoint's value lies precisely in the fact that anyone can reshape it into something else. Parameter-efficient fine-tuning (Hu et al., 2022) makes that reshaping inexpensive, so a downloader can adapt a released checkpoint to a new task at a small fraction of the cost of training from scratch. The same route lets a downloader push the model toward a harmful capability its publisher chose not to distribute: a text-to-image model can be fine-tuned until it reproduces an unauthorized artist's style again (Shan et al., 2023; Pham et al., 2024), and a safety-aligned language model can be fine-tuned until it exhibits a malicious behavior it was trained to refuse (Qi et al., 2024; Huang et al., 2026). Fig. 1 shows that a Stable Diffusion model fine-tuned on a few reference paintings by Vincent van Gogh can mimic his painting style (the image labeled with "w/o immunization").

Once a checkpoint is public, its weights cannot be withdrawn, and the publisher can neither revise them nor intervene at inference time; whatever protection exists must already be built into the released copy, before the publisher loses control of it. In response, model immunization intervenes at exactly this point: the owner modifies the weights themselves, rather than filtering the training data (O'Brien et al., 2025) or interposing at inference time (Schramowski et al., 2023; Inan et al., 2023), aiming to keep the checkpoint openly released while making one designated capability difficult to relearn. Images labeled with "Defense" in Fig. 1 show that, after the model is immunized by either IMMA or CLAMP, existing fine-tuning fails to produce a model that can mimic Vincent van Gogh's painting style.

Turning that difficulty into an optimizable objective requires the publisher to first specify and simulate an attacker, then optimize the released weights against it. Existing methods follow this recipe and together span diffusion, classification, and language models (Henderson et al., 2023; Zheng & Yeh, 2024; Zheng et al., 2025; Abdalla et al., 2025; Sarker et al., 2026). However, the defense's reported protection is optimized against a simulated attacker, and running that optimization requires pinning the attacker to one specific choice of the harmful data it trains on, the adaptation strategy

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.

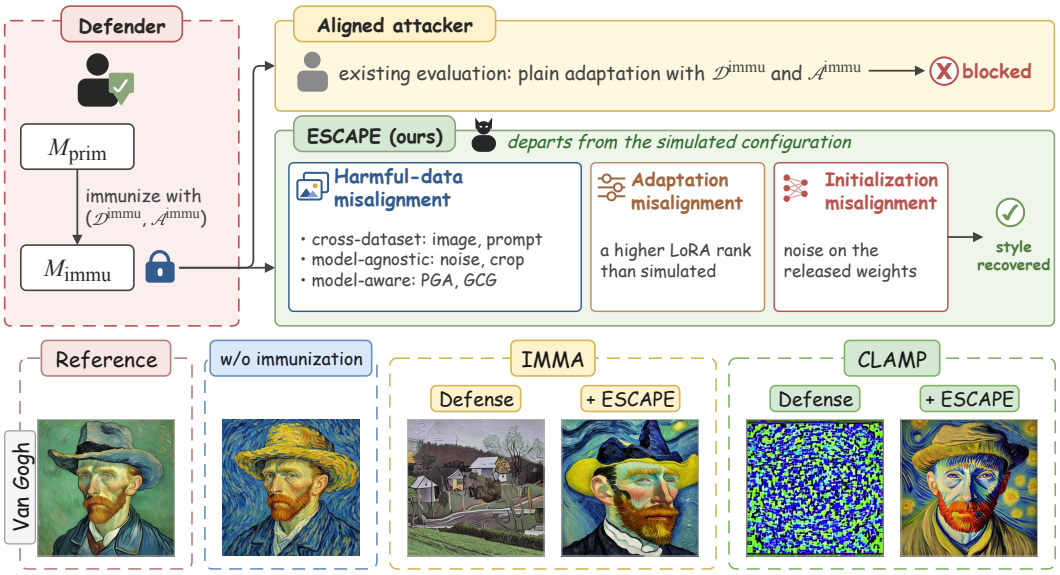

Figure 1: Overview of ESCAPE. Top: the defender immunizes M_{prim} into M_{immu} against a simulated attacker. An aligned attacker that reuses this configuration is blocked, whereas ESCAPE departs from it through harmful-data, adaptation, and initialization misalignment. Bottom: from left to right, a painting by Van Gogh, an image generated by a model adapted from an unimmunized model, plain adaptation of the model immunized by IMMA or CLAMP (Defense) misses the reference style, whereas ESCAPE recovers it.

it uses, and the model it starts adapting from. A real attacker is under no obligation to match any of these choices, so the protection such an evaluation reports may not generalize to unseen attacker configurations. In light of this, we propose ESCAPE, a framework that exploits the misalignment between the attacker simulated during immunization and the actual attacker after model release to expose vulnerabilities in existing immunization methods. As illustrated in Fig. 1, ESCAPE can circumvent existing immunizations and enable an immunized model to relearn Vincent van Gogh's painting style.

More specifically, ESCAPE systematically exploits the three attacker-controlled components fixed by existing immunization procedures. For harmful data, we replace the defender's proxy dataset with independently collected samples and further introduce model-agnostic or model-aware perturbations that alter the fine-tuning trajectory while preserving the targeted harmful capability. For adaptation, we vary the fine-tuning configuration, such as the LoRA rank, to access update directions not anticipated during immunization. For model initialization, we perturb the released checkpoint through transformations such as weight noise before fine-tuning. These simple escapes expose whether an immunization protects against the underlying harmful capability or merely against the particular attacker configuration used during its optimization.

We evaluate IMMA (Zheng & Yeh, 2024), CLAMP (Sarker et al., 2026), and CONDNUM (Zheng et al., 2025) across text-to-image generation, classification, and language modeling, spanning five artist styles, five classification benchmarks, and five language-model base models. Our results show that ESCAPE substantially weakens existing immunizations across tasks and modalities. In some settings, it completely eliminates the protection: the immunized model relearns the harmful capability at least as quickly as the corresponding unimmunized model.

Our contributions are summarized as follows:

- We expose the immunization–attacker misalignment overlooked by existing evaluations.
- We propose ESCAPE, which exploits misalignment in harmful data, adaptation strategy, and model initialization. Source code will be released upon acceptance.
- We evaluate across three modalities and three immunizations, showing that ESCAPE substantially weakens, and sometimes completely removes, existing immunization protection.

2 RELATED WORK

Model immunization. IMMA (Zheng & Yeh, 2024) immunizes a text-to-image diffusion model (Rombach et al., 2022) against a simulated attacker adapting it via low-rank adapters or DreamBooth (Hu et al., 2022; Ruiz et al., 2023), by moving the released weights to raise that adaptation's loss. This bilevel displacement starts from a concept-erased checkpoint (Gandikota et al., 2023). CONDNUM (Zheng et al., 2025) instead raises the condition number of the harmful task's feature Gram matrix on a classification backbone (He et al., 2016), slowing first-order optimization of a linear probe there while keeping the pretraining task well-conditioned. CLAMP (Sarker et al., 2026) instead reshapes the optimization dynamics of gradient-based fine-tuning so that long-horizon adaptation toward the harmful task stays contractive. This mechanism is not tied to one modality, so the same procedure applies to a diffusion model and to a safety-aligned language model. The three sit in a wider release-time line that starts with task blocking, which meta-learns weights whose adaptation to a harmful task fails (Henderson et al., 2023). Further work in this line immunizes several concepts at once (Zheng & Yeh, 2025), immunizes without committing to the attacker's adaptation algorithm (Abdalla et al., 2025), obstructs few-shot classification over restricted classes (Zheng et al., 2024), makes a checkpoint non-fine-tunable on restricted tasks (Deng et al., 2024), locks an encoder against probing on prohibited domains (Ding et al., 2025), and states immunization conditions for language models (Rosati et al., 2024; Huang et al., 2026). IMMA, CONDNUM, and CLAMP each report an immunized-to-un-immunized ratio for the attacker they specify and optimize their released weights against. We reproduce IMMA and CONDNUM and treat the protection each reports as valid within its original setting. We then evaluate all three methods under a common protocol to measure what the reported protection depends on.

Capability removal and protective perturbations. Immunization usually sits atop capability removal: weight-level concept erasure (Gandikota et al., 2023; 2024), inference-time steering (Schramowski et al., 2023), or classifier unlearning (Fan et al., 2024). Fine-tuning is known to reverse erasure alone (Pham et al., 2024; Zheng & Yeh, 2024), which motivates immunization. We hold that checkpoint fixed and vary only the immunization displacement. A parallel line instead has the *defender* perturb the *images*, so models trained on them cannot acquire the capability: unlearnable examples (Huang et al., 2021), style-mimicry protection (Shan et al., 2023), anti-personalization perturbations (Le et al., 2023), and edit-resistant images (Salman et al., 2023). Here the *attacker* perturbs its own images to promote learning, yet diagnostics transfer: availability attacks are known to install a shortcut while leaving the underlying information intact (Yu et al., 2022). We interpret a perturbation that helps only on the immunized arm the same way.

3 THREAT MODEL

Defender's capabilities. The defender constructs the immunized model M_{immu} from an open-weight model M_{prim} trained on the primary-task dataset $\mathcal{D}_{\text{prim}}$, and therefore has white-box access to the model throughout the immunization process. The defender has access to $\mathcal{D}_{\text{prim}}$ and controls the harmful proxy dataset $\mathcal{D}_{\text{harm}}^{\text{immu}}$ and all configurations of the immunization procedure. The defender may also simulate known adaptation attacks $\mathcal{A}^{\text{immu}}$ during immunization. Once M_{immu} is released, however, the defender cannot control how downstream users access, modify, or adapt the model. In particular, the defender does not know the attacker's adaptation dataset $\mathcal{D}_{\text{harm}}^{\text{attk}}$ or adaptation strategy $\mathcal{A}^{\text{attk}}$, all of which may differ from those simulated during immunization.

Defender's goals. An effective immunized model should satisfy the following three requirements: 1) *Primary Utility:* The immunized model M_{immu} should preserve the original model's performance on the primary task, as measured on $\mathcal{D}_{\text{prim}}$. 2) *Harmful Inadaptability:* An attacker should be unable to derive a useful harmful model from M_{immu} using any feasible adaptation strategy. Formally, for an attacker-selected adaptation strategy $\mathcal{A}^{\text{attk}}$ and harmful dataset $\mathcal{D}_{\text{harm}}^{\text{attk}}$, the resulting model $M_{\text{harm}} = \mathcal{A}^{\text{attk}}(M_{\text{immu}}, \mathcal{D}_{\text{harm}}^{\text{attk}})$ should exhibit low performance on the designated harmful capability. 3) *Benign Adaptability:* A benign user should remain able to adapt M_{immu} to tasks unrelated to the designated harmful capability. Specifically, given a benign adaptation strategy $\mathcal{A}^{\text{beni}}$ and dataset $\mathcal{D}_{\text{beni}}$, the resulting model $M_{\text{beni}} = \mathcal{A}^{\text{beni}}(M_{\text{immu}}, \mathcal{D}_{\text{beni}})$ should achieve high performance on the corresponding benign task, ideally comparable to a model adapted from M_{prim}.

Attacker's capabilities. The attacker obtains the released immunized model M_{immu} and has white-box access to it. The attacker knows from failed adaptation attempts that the model has been immunized. However, ESCAPE is immunization-agnostic: it requires no knowledge of the immunization method. The attacker controls both the harmful adaptation dataset $\mathcal{D}_{\text{harm}}^{\text{attk}}$ and the complete adaptation strategy $\mathcal{A}^{\text{attk}}$, and may also perturb the released weights before adapting them. The attacker may also transform $\mathcal{D}_{\text{harm}}^{\text{attk}}$ while preserving its harmful objective. Such transformations may be model-agnostic, such as additive random noise, or model-aware, such as gradient-guided perturbations.

Attacker's goal. The attacker's goal is to obtain a model $M_{\text{harm}} = \mathcal{A}^{\text{attk}}(M_{\text{immu}}, \mathcal{D}_{\text{harm}}^{\text{attk}})$ that has a similar performance as $\mathcal{A}^{\text{attk}}(M_{\text{prim}}, \mathcal{D}_{\text{harm}}^{\text{attk}})$ on the designated harmful capability.

4 OUR IMMUNE ESCAPE

As described in our threat model, the defender approximates the unknown post-release attacker during immunization using a harmful proxy dataset $\mathcal{D}_{\text{harm}}^{\text{immu}}$ and a surrogate adaptation strategy $\mathcal{A}^{\text{immu}}$. A key limitation of existing evaluations is that they typically reuse this same dataset and adaptation strategy to measure immunization effectiveness. Such an aligned evaluation demonstrates resistance only to the specific attack configuration anticipated during immunization and may therefore obscure configuration-specific overfitting. After M_{immu} is released, the defender cannot control either the data or the adaptation strategy selected by a downstream attacker. Consequently, there is no reason to assume that $\mathcal{D}_{\text{harm}}^{\text{attk}} = \mathcal{D}_{\text{harm}}^{\text{immu}}$, $\mathcal{A}^{\text{attk}} = \mathcal{A}^{\text{immu}}$, or M_{immu} stays unperturbed before adaptation. ESCAPE exploits this misalignment along each axis the attacker controls.

4.1 ESCAPE VIA HARMFUL-DATA MISALIGNMENT

We introduce three types of harmful-data misalignment: 1) cross-dataset misalignment, 2) model-agnostic perturbation, and 3) model-aware perturbation.

Cross-dataset misalignment. The harmful capability targeted by immunization corresponds to an underlying concept or data distribution rather than to a particular finite dataset. Because the defender cannot enumerate this distribution directly, they collect a finite proxy dataset $\mathcal{D}_{\text{harm}}^{\text{immu}} = \{(x_i, y_i)\}_{i=1}^{n} \sim \mathcal{P}_{\text{harm}}$ and use it to simulate harmful adaptation during immunization. This dataset inevitably provides only a limited empirical representation of the harmful capability.

Existing evaluations commonly assess immunization using the same harmful samples employed during immunization. Although this aligned setting measures resistance on $\mathcal{D}_{\text{harm}}^{\text{immu}}$, it does not establish that the learned protection generalizes to other samples representing the same harmful capability. In practice, a downstream attacker can independently collect another dataset $\mathcal{D}_{\text{harm}}^{\text{attk}} = \{(x'_j, y'_j)\}_{j=1}^{m} \sim \mathcal{P}_{\text{harm}}$, but $\mathcal{D}_{\text{harm}}^{\text{attk}} \neq \mathcal{D}_{\text{harm}}^{\text{immu}}$. The two datasets target the same harmful capability but contain different samples and may exhibit different low-level characteristics.

To exploit this cross-dataset misalignment, ESCAPE replaces the defender's proxy dataset with an independently collected harmful dataset while keeping the adaptation strategy fixed. Specifically, we construct $M_{\text{harm}} = \mathcal{A}^{\text{attk}}(M_{\text{immu}}, \mathcal{D}_{\text{harm}}^{\text{attk}})$, where $\mathcal{D}_{\text{harm}}^{\text{attk}}$ represents the same harmful capability as $\mathcal{D}_{\text{harm}}^{\text{immu}}$ but does not reuse its training samples.

This attack naturally applies across modalities. For image classifiers, the attacker can use different images representing the same harmful target classes. For language models, alternative instruction–response pairs can instantiate the same harmful capability. For text-to-image models, alternative image–text pairs can represent the same prohibited concept. Such multi-modal models further expose multiple sources of misalignment: the attacker can alter the visual component while preserving the textual component (image misalignment), alter the textual component while preserving the visual component (prompt misalignment), or change both components jointly. App. E explains how we achieve textual misalignment for language and text-to-image models.

Model-agnostic perturbation on images. Without accessing M_{immu}, the attacker can apply standard image transformations. We consider two transformations: random noise $\delta \sim \mathcal{U}(-\varepsilon, \varepsilon)$, followed by clipping $x + \delta$ to the valid pixel range, and cropping followed by rescaling (App. A). On

classification, the random noise instead sets every pixel to $\pm\varepsilon$ with a random sign, matching the amplitude of the model-aware perturbation, as App. A details. These transformations preserve the harmful semantic content while altering low-level image statistics. Because they require no model weights, gradients, outputs, or knowledge of the immunization method, they represent a minimal-capability attack baseline.

Model-aware perturbation. Existing immunizations commonly modify M_{immu} such that harmful fine-tuning on $\mathcal{D}_{\text{harm}}^{\text{immu}}$ produces misleading parameter gradients, such as pointing to the opposite direction from successful harmful adaptation. We exploit this distorted gradient signal to perturb $\mathcal{D}_{\text{harm}}^{\text{attk}}$ so that fine-tuning M_{immu} on the perturbed dataset induces gradients that are better aligned with a productive harmful-adaptation direction. In continuous domains such as images, we apply PGA, while in discrete domains such as text, we search along the direction guided by the largest gradient. In both cases, the released weights stay fixed and only the attacker's input is optimized. We next describe how we instantiate model-aware perturbation for diffusion and language models. Due to space constraints, we defer the corresponding instantiation for image classifiers to App. C.

✦ *Perturbing images in diffusion models.* For diffusion models, immunization modifies the local loss landscape around M_{immu} such that harmful fine-tuning on a sample $x \sim \mathcal{D}_{\text{harm}}^{\text{attk}}$ follows an ineffective optimization trajectory within an immunization-induced basin. To escape this basin, we construct a bounded perturbation δ that maximizes the diffusion model's noise-prediction loss: $\delta^* = \arg\max_{\|\delta\|_\infty \le \varepsilon} \mathcal{L}_{\text{diff}}(M_{\text{immu}}; x+\delta)$, where ε bounds the perturbation magnitude. The resulting sample $x+\delta^*$ preserves the harmful concept represented by x but induces a larger prediction error in M_{immu}. Fine-tuning on the perturbed sample therefore produces a stronger and qualitatively different gradient signal, pulling the optimization trajectory away from the immunization-induced basin and enabling the model to relearn the harmful concept.

✦ *Perturbing text in language models.* Unlike images, whose pixels can be optimized directly in a continuous space, text consists of discrete tokens. Given a harmful instruction–response pair $(q, a) \sim \mathcal{D}_{\text{harm}}^{\text{attk}}$, we therefore augment the instruction q with an optimizable suffix s. The objective is to identify a suffix that induces M_{immu} to assign a high likelihood to the harmful response a conditioned on the augmented instruction $q \oplus s$: $s^* = \arg\min_{s \in \mathcal{V}^L} [-\lambda \log(1 - P_{\text{refusal}}) - \log P_{M_{\text{immu}}}(a \mid q \oplus s)]$, where $\mathcal{V}$ is the token vocabulary, L is the suffix length, and P_{refusal} denotes the probability that M_{immu} refuses to answer, which App. C scores at the first generated token. We write $\log P_{M_{\text{immu}}}(a \mid q \oplus s)$ for the average over the tokens of a, which is truncated to a fixed length, so that this term and the refusal term stay on a comparable scale. Because direct optimization over discrete tokens is infeasible, we use gradients with respect to the token representations to guide the search for s^* like GCG (Zou et al., 2023). We then replace (q, a) with $(q \oplus s^*, a)$ during harmful fine-tuning.

4.2 Other escapes

Adaptation misalignment. For diffusion and language models, both immunization and the simulated attacker $\mathcal{A}^{\text{immu}}$ used during immunization typically rely on low-rank adaptation (LoRA), a widely adopted parameter-efficient fine-tuning method. However, the adaptation configuration assumed during immunization need not match the one used by the actual attacker. We instantiate this misalignment by varying the LoRA rank. For example, a model immunized against rank-4 LoRA may subsequently be adapted using rank-16 LoRA. The higher-rank adapter spans a larger update space, potentially exposing optimization directions that were insufficiently constrained during immunization.

Model initialization misalignment. Immunization methods typically assume that the attacker directly adapts the immunized model M_{immu}. In practice, however, the attacker may first perturb the model parameters before adaptation, causing optimization to start from a different initialization. We instantiate this misalignment using common model transformations such as weight quantization and the injection of random noise. Such perturbations may alter the subsequent optimization trajectory and allow adaptation to escape the protection established around M_{immu}.

5 EXPERIMENTS

5.1 SETUP

Immunizations and models. We evaluate three immunizations using their original configurations. IMMA (Zheng & Yeh, 2024) is evaluated on Stable Diffusion v1.4 (Rombach et al., 2022), COND-NUM (Zheng et al., 2025) on a ResNet-18 classifier (He et al., 2016), and CLAMP (Sarker et al., 2026) on both architectures as well as five language models: Llama-3.2-1B (Grattafiori et al., 2024), Gemma-2-2B (Gemma Team, 2024), Qwen2-7B (Yang et al., 2024), Qwen2.5-1.5B (Yang et al., 2025a), and Qwen3-1.7B (Yang et al., 2025b). For each immunization, we use either the official implementation or the checkpoints provided by the authors.

Datasets. The diffusion experiments cover five artist styles: Van Gogh, Tyler Edlin, Kilian Eng, Thomas Kinkade, and Kelly McKernan. The classification experiments cover five harmful classification tasks: Stanford Cars (Krause et al., 2013), a 196-category fine-grained car-model recognition task; DTD (Cimpoi et al., 2014), a 47-category surface-texture recognition task; Food-101 (Bossard et al., 2014), a 101-category dish recognition task; Flowers-102 (Nilsback & Zisserman, 2008), a 102-category flower recognition task; and VMMRdb-196 (Tafazzoli et al., 2017), a 196-category vehicle make-and-model recognition task drawn from photos collected on the web. The language-model experiments use three disjoint splits of SG-Bench (Mou et al., 2024): 596 examples for constructing the defense, 238 for optimizing the attack, and 358 held out for scoring.

Implementation. Unless otherwise specified, we follow the adaptation strategies used in the original immunization studies, which we refer to as *plain adaptation*. For PGA on images, we constrain the perturbation to $\|\delta\|_\infty \leq 1/255$ for image classification and $\|\delta\|_\infty \leq 4/255$ for diffusion models. We use the same perturbation budgets for the corresponding model-agnostic controls. For GCG on language models, we optimize a 20-token suffix. Additional details are provided in App. A and C.

5.2 EVALUATION METRICS

We use the same metrics defined in the literature to evaluate the immunization effects across different model families. Similarity Gap Ratio (SGR) was proposed by IMMA (Zheng & Yeh, 2024) to measure the immunization gain on diffusion models and later extended to classification and language models (Sarker et al., 2026). The general definition of SGR is:

$$\mathrm{SGR}(M_{\text{prim}}, M_{\text{immu}}) = \frac{U(M_{\text{prim}}) - U(M_{\text{immu}})}{U(M_{\text{prim}})} \times 100, \tag{1}$$

where $U(M)$ measures the adaptability of a model M, that is, the harmful-task performance of a model adapted from M. Intuitively, an effective immunized model should have a lower adaptability and thus a high SGR. In contrast, the attacker desires a lower SGR. A lower SGR means that less protection remains, while a negative value means that the immunized model achieves higher harmful-task adaptability than the original model.

For *diffusion models*, we measure adaptability by the similarity between images generated by the adapted model and the reference. Following IMMA, we use the average similarities computed by CLIP (Radford et al., 2021), DINO (Caron et al., 2021), and $1 - \mathrm{LPIPS}$ (Zhang et al., 2018). For *classification models*, we measure adaptability by classification accuracy. We further report ACC_O and ACC_I that denote the accuracy of the model adapted from M_{prim} and M_{immu}, respectively. For *language models*, we measure adaptability via harmful response rate (HR) judged by Llama-Guard-3-1B (Inan et al., 2023; Grattafiori et al., 2024). Similarly, we report HR_O and HR_I.

5.3 RESULTS ON DIFFUSION MODELS

Harmful-data misalignment. Tab. 1 reports each harmful-data misalignment of Sec. 4.1 and their combination, where $\pm$ denotes the standard deviation over three seeds. Cross-dataset misalignment, which replaces the images (Image misalignment) or rewrites only the prompt (Prompt misalignment), model-agnostic noise, and model-aware PGA each lower SGR relative to plain adaptation in every defense–artist cell, although a few single-component reductions fall within one standard

Table 1: Main generation results. Average SGR; lower is better for the attacker. Combined applies image misalignment, prompt misalignment, and model-aware PGA together. The model-agnostic row is the matched control for model-aware PGA. Bold marks the best attack for each defense and artist.

Defense	Attack	Van Gogh	T. Edlin	K. Eng	T. Kinkade	K. McKernan
IMMA	Plain adaptation	$17.68_{\pm 5.45}$	$20.26_{\pm 0.84}$	$21.98_{\pm 5.92}$	$19.65_{\pm 0.81}$	$24.78_{\pm 0.58}$
	Image misalignment	$17.07_{\pm 1.75}$	$17.13_{\pm 2.98}$	$10.28_{\pm 5.19}$	$11.35_{\pm 2.65}$	$12.15_{\pm 2.72}$
	Prompt misalignment	$\mathbf{6.46}_{\pm 1.20}$	$\mathbf{8.98}_{\pm 1.40}$	$11.84_{\pm 1.99}$	$9.13_{\pm 2.40}$	$9.26_{\pm 2.78}$
	Model-aware (PGA)	$8.41_{\pm 0.33}$	$13.93_{\pm 0.37}$	$11.15_{\pm 2.92}$	$13.07_{\pm 2.59}$	$16.27_{\pm 0.42}$
	Model-agnostic (noise)	$14.08_{\pm 0.41}$	$14.82_{\pm 1.75}$	$13.50_{\pm 2.37}$	$17.01_{\pm 2.35}$	$21.95_{\pm 3.63}$
	Combined	$\mathbf{6.46}_{\pm 2.07}$	$12.91_{\pm 1.65}$	$\mathbf{6.80}_{\pm 1.44}$	$\mathbf{6.23}_{\pm 3.16}$	$\mathbf{-1.53}_{\pm 1.57}$
CLAMP	Plain adaptation	$18.41_{\pm 1.09}$	$40.97_{\pm 0.05}$	$25.36_{\pm 0.40}$	$32.12_{\pm 1.75}$	$38.81_{\pm 0.12}$
	Image misalignment	$12.78_{\pm 3.35}$	$32.40_{\pm 1.15}$	$11.34_{\pm 0.80}$	$19.96_{\pm 2.15}$	$\mathbf{29.88}_{\pm 7.55}$
	Prompt misalignment	$7.88_{\pm 2.64}$	$38.65_{\pm 1.85}$	$16.56_{\pm 3.69}$	$27.41_{\pm 10.15}$	$32.91_{\pm 3.43}$
	Model-aware (PGA)	$10.05_{\pm 0.60}$	$35.06_{\pm 3.96}$	$18.24_{\pm 1.80}$	$21.74_{\pm 2.41}$	$34.34_{\pm 2.72}$
	Model-agnostic (noise)	$14.56_{\pm 2.40}$	$38.49_{\pm 0.77}$	$20.24_{\pm 3.92}$	$24.33_{\pm 2.59}$	$34.01_{\pm 2.77}$
	Combined	$\mathbf{7.07}_{\pm 0.87}$	$\mathbf{17.93}_{\pm 2.30}$	$\mathbf{6.40}_{\pm 4.30}$	$\mathbf{14.23}_{\pm 3.81}$	$31.92_{\pm 2.87}$

Table 2: Change in style similarity from plain adaptation to model-aware PGA, averaged over the three metrics. ΔU_O and ΔU_I are the changes in the similarity reached after adapting M_{prim} and M_{immu}. App. F gives the full results.

Defense		Van Gogh	T. Edlin	K. Eng	T. Kinkade	K. McKernan
IMMA	ΔU_O	$+0.001$	$+0.007$	$+0.003$	-0.009	-0.006
	ΔU_I	$+0.053$	$+0.047$	$+0.070$	$+0.039$	$+0.053$
CLAMP	ΔU_O	-0.003	-0.005	$+0.003$	$+0.005$	-0.025
	ΔU_I	$+0.045$	$+0.039$	$+0.046$	$+0.073$	$+0.017$

deviation. Combining them reaches the lowest SGR in most cells; on IMMA with Kelly McKernan it lowers SGR from 24.78 to -1.53. Under model-aware PGA, the similarity reached by adapting M_{immu} rises by between 0.017 and 0.073, while that reached by adapting M_{prim} changes by at most 0.025, so the reduction comes from M_{immu} reaching a higher similarity. Tab. 2 reports the change for each defense and artist. Fig. 2 shows some qualitative results, demonstrating our method can achieve immune escape. More results are provided in App. F.

Adaptation misalignment. The attacker changes the LoRA rank r of $\mathcal{A}^{\text{attk}}$ from 4 to 8, 16, and 32. See Sec. 4.2 for the adaptation attack. Tab. 3 shows that SGR falls as r grows on both defenses and both artists. At $r = 32$ it drops from 24.78 to 13.17 on IMMA with Kelly McKernan and from 38.81 to 29.69 on CLAMP with Kelly McKernan.

Table 3: LoRA-rank misalignment on diffusion (SGR); plain adaptation uses r=4.

Defense	Attack	Van Gogh	K. McKernan
IMMA	Plain adapt.	$17.68_{\pm 5.45}$	$24.78_{\pm 0.58}$
	$r = 8$	$14.76_{\pm 4.11}$	$17.99_{\pm 3.92}$
	$r = 16$	$10.92_{\pm 3.78}$	$17.40_{\pm 2.95}$
	$r = 32$	$\mathbf{8.72}_{\pm 2.26}$	$\mathbf{13.17}_{\pm 3.37}$
CLAMP	Plain adapt.	$18.41_{\pm 1.09}$	$38.81_{\pm 0.12}$
	$r = 8$	$18.20_{\pm 5.16}$	$36.75_{\pm 1.46}$
	$r = 16$	$14.59_{\pm 2.24}$	$34.57_{\pm 2.46}$
	$r = 32$	$\mathbf{11.01}_{\pm 2.86}$	$\mathbf{29.69}_{\pm 1.73}$

Model initialization misalignment. Perturbing the weights of M_{immu} with small Gaussian noise before adaptation also weakens immunization: against CLAMP on Van Gogh, every noise level we test lowers SGR from 18.41 to between 7.95 and 10.38. App. A reports the results for each noise level.

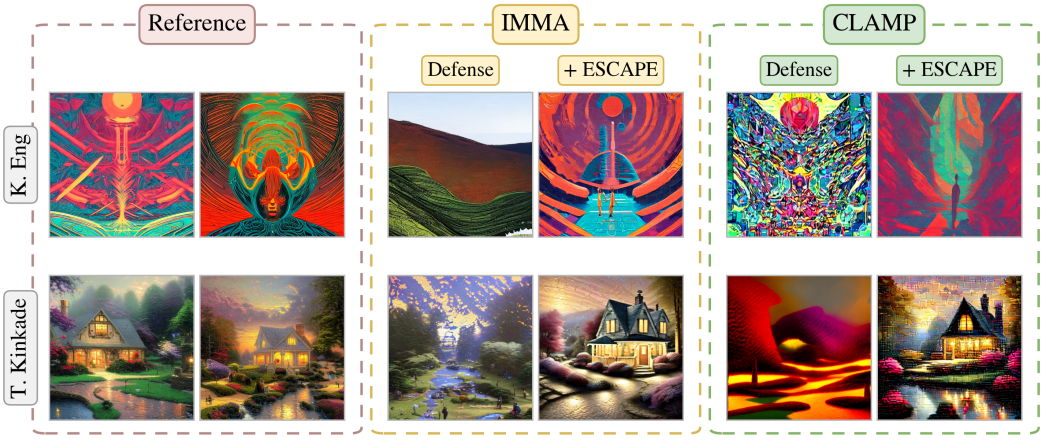

Figure 2: Adapted samples on two protected styles. *Defense* adapts M_{immu} with plain adaptation, and +ESCAPE with the Combined attack (Tab. 1). The defense-only samples miss the reference style and the ESCAPE samples recover it, which is what the SGR reductions in Tab. 1 measure.

Table 4: CONDNUM classification results. For SGR, lower is better for the attacker. ACC_I and ACC_O (%) are the probe accuracies after adapting M_{immu} and M_{prim}, averaged over epochs 10/20/30/40/50. ΔACC_I is the relative change of ACC_I from plain adaptation. App. F reports ACC_O under the two attacks.

	Plain adaptation			Model-agnostic (noise)			Model-aware (PGA)		
Dataset	SGR	ACC_I	ACC_O	SGR	ACC_I	ΔACC_I	SGR	ACC_I	ΔACC_I
Stanford Cars	$20.99_{\pm 5.90}$	$22.67_{\pm 1.38}$	$28.28_{\pm 0.32}$	7.07	$26.28_{\pm 0.17}$	+15.92%	**−0.25**	$28.35_{\pm 0.52}$	+25.06%
DTD	$7.13_{\pm 0.59}$	$57.91_{\pm 0.52}$	$62.34_{\pm 0.30}$	6.40	$58.35_{\pm 0.33}$	+0.76%	**5.69**	$58.79_{\pm 0.16}$	+1.52%
Food-101	$4.65_{\pm 1.68}$	$45.52_{\pm 0.75}$	$47.67_{\pm 0.25}$	3.25	$46.12_{\pm 0.26}$	+1.32%	**2.33**	$46.56_{\pm 0.28}$	+2.28%
Flowers-102	$5.61_{\pm 0.99}$	$71.15_{\pm 0.53}$	$75.16_{\pm 0.33}$	6.33	$70.40_{\pm 0.41}$	−1.05%	**−0.59**	$75.60_{\pm 0.29}$	+6.25%
VMMRdb-196	$35.58_{\pm 3.01}$	$7.78_{\pm 0.50}$	$11.72_{\pm 0.39}$	29.61	$8.25_{\pm 0.33}$	+6.04%	**19.45**	$9.44_{\pm 0.46}$	+21.34%

5.4 RESULTS ON CLASSIFICATION MODELS

See App. F for the full results. App. C describes the frozen CONDNUM feature extractor. We apply PGA against it and train the same linear probe on the perturbed features. Model-agnostic noise at the same budget serves as the control. Tab. 4 reports that model-aware PGA raises ACC_I above both plain adaptation and the model-agnostic control on every dataset. On some datasets, ACC_I with PGA is higher than ACC_O.

5.5 RESULTS ON LANGUAGE MODELS

We evaluate CLAMP on the five language models and SG-Bench splits that Sec. 5.1 describes. See App. E for the prompt-misalignment results. The three arms differ only in the suffix appended to the requests in $\mathcal{D}^{\text{attk}}_{\text{harm}}$: none (plain adaptation), a GCG-optimized suffix, or the best of N random suffixes. Tab. 5 shows that under plain adaptation, HR_I stays between 0.56 and 4.13, far below HR_O, which ranges from 3.48 to 27.37. The GCG suffix raises HR_I to between 59.84 and 73.53. GCG-constructed fine-tuning data makes harmful fine-tuning substantially more effective for both the original and the immunized model, and largely erases CLAMP's relative protection. App. F shows the same with a rank-4 adapter.

5.6 ABLATION STUDIES AND ANALYSES

Model-aware versus model-agnostic perturbation. Each model-aware perturbation is compared with a matched control at the same budget, isolating how much of the effect requires the gradients of M_{immu}. Tab. 1 reports that, on diffusion models, model-aware PGA reaches a lower SGR than model-agnostic noise in nine of the ten defense–artist cells; the exception is Kelly McKernan under CLAMP (34.34 vs. 34.01). Tab. 4 shows that, on classification, it does so on all five datasets and

Table 5: CLAMP language-model results. For SGR, lower is better for the attacker. HR_I and HR_O (%) are the harmful rates after adapting M_{immu} and M_{prim}. $\Delta\mathrm{HR}_I$ is the relative change of HR_I from plain adaptation. App. F reports HR_O under the two attacks.

	Plain adaptation			Random suffix			Model-aware (GCG)		
Model	SGR	HR_I	HR_O	SGR	HR_I	$\Delta\mathrm{HR}_I$	SGR	HR_I	$\Delta\mathrm{HR}_I$
Llama-3.2-1B	71.32	$2.67_{\pm 0.05}$	$9.31_{\pm 0.57}$	-170.35	$25.17_{\pm 0.19}$	$+842.70\%$	$\mathbf{-615.36}$	$66.60_{\pm 0.23}$	$+2394.38\%$
Gemma-2-2B	83.91	$0.56_{\pm 0.00}$	$3.48_{\pm 0.51}$	-50.86	$5.25_{\pm 0.28}$	$+837.50\%$	$\mathbf{-1760.34}$	$64.74_{\pm 1.59}$	$+11460.71\%$
Qwen2.5-1.5B	80.50	$1.24_{\pm 0.11}$	$6.36_{\pm 0.55}$	-125.47	$14.34_{\pm 0.81}$	$+1056.45\%$	$\mathbf{-840.88}$	$59.84_{\pm 0.86}$	$+4725.81\%$
Qwen2-7B	44.16	$3.97_{\pm 0.30}$	$7.11_{\pm 0.61}$	-522.08	$44.23_{\pm 1.68}$	$+1014.11\%$	$\mathbf{-820.25}$	$65.43_{\pm 0.30}$	$+1548.11\%$
Qwen3-1.7B	84.91	$4.13_{\pm 0.11}$	$27.37_{\pm 0.76}$	44.65	$15.15_{\pm 0.33}$	$+266.83\%$	$\mathbf{-168.65}$	$73.53_{\pm 0.70}$	$+1680.39\%$

Table 6: Defense-training distribution generalization on CONDNUM. *Real* immunizes on the real training set; *FLUX* immunizes on FLUX.2 synthetic images. The attacker probes on the real training set and every row is scored on the real test set. Accuracy is measured at epoch 50, while SGR averages epochs $10/20/30/40/50$.

Dataset	$\mathcal{D}_{\text{harm}}^{\text{immu}}$	$\mathrm{ACC}_I@50$	$\mathrm{ACC}_O@50$	SGR
Stanford Cars	Real	$28.08_{\pm 1.85}$	$32.95_{\pm 0.37}$	$20.99_{\pm 5.90}$
	FLUX	$32.46_{\pm 0.15}$	$33.15_{\pm 0.28}$	$7.46_{\pm 1.41}$
DTD	Real	$59.50_{\pm 0.59}$	$63.21_{\pm 0.48}$	$7.13_{\pm 0.59}$
	FLUX	$59.77_{\pm 0.13}$	$63.17_{\pm 0.62}$	$4.94_{\pm 0.82}$
Flowers-102	Real	$78.73_{\pm 0.68}$	$78.11_{\pm 0.33}$	$5.61_{\pm 0.99}$
	FLUX	$80.06_{\pm 0.11}$	$78.02_{\pm 0.38}$	$-4.90_{\pm 0.91}$
VMMRdb-196	Real	$10.81_{\pm 0.80}$	$14.70_{\pm 0.61}$	$35.58_{\pm 3.01}$
	FLUX	$15.41_{\pm 0.49}$	$14.25_{\pm 0.28}$	$7.45_{\pm 1.07}$

drives SGR below zero on Stanford Cars and Flowers-102, where the control remains positive. Tab. 5 compares this on language models, where the GCG suffix exceeds the best of N random suffixes, drawn at GCG's own candidate budget and scored by its objective, by between 21.20 and 59.49 points in HR_I on every model. See App. F for a sweep of the PGA step size used by the Combined attack on generation.

Defense-training distribution generalization on classification. In practice the defender does not have access to the real harmful training set and instead approximates $\mathcal{D}_{\text{harm}}^{\text{immu}}$ from the class names alone, using images synthesized by FLUX.2 [klein] 9B (Black Forest Labs, 2026). On CONDNUM, these synthetic images sit on the defender's side rather than the attacker's, so this analysis changes the defender's training data and is not an instance of image misalignment. The defender generates $\mathcal{D}_{\text{harm}}^{\text{immu}}$ this way and runs CONDNUM on it, keeping ImageNet (Deng et al., 2009) as the benign set. The attacker probes the immunized model on the real training set exactly as in App. C. The reference row is immunized on real images. The original extractor M_{prim}, the attacker's data, and the real test set are shared, so the data the defender immunized on is the only variable. Tab. 6 reports that the synthetic $\mathcal{D}_{\text{harm}}^{\text{immu}}$ lowers SGR on all four datasets, and lowers it below zero on Flowers-102, while $\mathrm{ACC}_O@50$ moves by at most 0.45 points.

6 CONCLUSION

Model immunization is optimized against a defender-simulated attacker, so its protection holds only near that configuration, and even an unoptimized perturbation to the attacker's data can weaken it. We address this with ESCAPE, which misaligns the attacker's harmful data, adaptation strategy, and model initialization from those the defender simulated. Its core, model-aware perturbation, perturbs the attacker's input along the released model's gradient, instantiated as PGA over continuous pixels and GCG over discrete tokens. We evaluate model-aware perturbation on text-to-image diffusion, classification, and language models, and it reduces the measured protection in every defense and setting we evaluate. Combining all three components on the generative modality, ESCAPE attains the lowest or tied-lowest measured protection in eight of ten defense–artist settings. A reported protection number is therefore interpretable only together with the attacker configuration against which it was measured.

ETHICS STATEMENT

This paper reports attacks against defenses that exist to protect artists' styles and to make released checkpoints resist unauthorized fine-tuning, so the reason for reporting them deserves to be stated directly. Three of the misalignments we measure need no access to the released model at all: image misalignment, prompt misalignment, and a model-agnostic pixel perturbation, each built only from means that were already available before IMMA and CLAMP were published. This paper measures how much protection survives an attacker who uses only those pre-existing means. A defender or an artist relying on the defense can then compare the number reported at release against the protection that actually remains once an unremarkable user departs from the exact attacker configuration the defense was tuned against. App. G turns this into concrete guidance for defenders.

Sec. 4.1 defines image misalignment: on the generative defense, it substitutes images that the attacker generates for itself, conditioned on the defender's own prompt, from a publicly available text-to-image checkpoint the defender never immunized against. No artwork is collected from an artist for this purpose. See App. D for the construction protocol. These substitute images serve only to measure how much protection survives that substitution.

Every model and dataset used in this paper is public. The generative experiments start from a Stable Diffusion v1.4 checkpoint with an artistic style erased by a published method, apply the IMMA and CLAMP procedures using each defense's own published weights and training code, adapt with low-rank adapters, and score similarity with LPIPS, CLIP, and DINO. The classification experiments immunize an ImageNet-pretrained ResNet-18 with CONDNUM, retain ImageNet (Deng et al., 2009) for the utility check, and probe the harmful task on Stanford Cars, DTD, Food-101, Flowers-102, and VMMRdb, all public benchmark datasets. None of this involves human subjects, personal data, or proprietary material. No checkpoint produced by an attack in this paper is deployed to a running or public-facing system; each exists only as an experimental artifact used to produce the measurements reported here. The authors declare no competing interests.

REPRODUCIBILITY STATEMENT

Sec. 3 states the threat model, and Sec. 4 defines the harmful-data, adaptation, and model initialization misalignments that ESCAPE exploits, and Sec. 4.1 covers cross-dataset misalignment (image and prompt misalignment), model-agnostic perturbation, and model-aware perturbation in turn. See App. G for the dimensions, including the attacker's adaptation strategy, a defense evaluation should state and vary independently. App. A and App. C collect implementation-level detail behind those constructions, perturbation budgets, optimizer settings, and the configuration specific to each of the three evaluated defenses. App. B lists the checks an attack must clear to be credited: a matched control, absolute utility beside every ratio, and two-arm specificity.

Sec. 5 reports image misalignment, prompt misalignment, model-agnostic perturbation, model-aware PGA and GCG, their Combined attack, adaptation misalignment, and model initialization misalignment. Apps. A–F give additional attack construction, evaluation, and implementation detail.

USE OF LARGE LANGUAGE MODELS

Large language models were used only to help edit and polish the writing in this paper. All technical content, including the threat model, the attack constructions, the experiments, and the conclusions drawn from them, is the authors' own work, and the authors take full responsibility for the content of this paper.

REFERENCES

Amro Abdalla, Ismail Shaheen, Dan DeGenaro, Rupayan Mallick, Bogdan Raita, and Sarah Adel Bargal. GIFT: gradient-aware immunization of diffusion models against malicious fine-tuning with safe concepts retention. *CoRR*, abs/2507.13598, 2025.

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pp. 284–293. PMLR, 2018.

Black Forest Labs. FLUX.2 [klein]: Towards interactive visual intelligence, 2026. URL https://bfl.ai/blog/flux2-klein-towards-interactive-visual-intelligence.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pp. 9630–9640. IEEE, 2021.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pp. 3606–3613. IEEE Computer Society, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255. IEEE Computer Society, 2009.

Jiangyi Deng, Shengyuan Pang, Yanjiao Chen, Liangming Xia, Yijie Bai, Haiqin Weng, and Wenyuan Xu. Sophon: Non-fine-tunable learning to restrain task transferability for pre-trained models. In *SP*, pp. 2553–2571. IEEE, 2024.

Ruyi Ding, Tong Zhou, Lili Su, Aidong Adam Ding, Xiaolin Xu, and Yunsi Fei. Probe-me-not: Protecting pre-trained encoders from malicious probing. In *NDSS*. The Internet Society, 2025.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *ICLR*. OpenReview.net, 2024.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *ICCV*, pp. 2426–2436. IEEE, 2023.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzynska, and David Bau. Unified concept editing in diffusion models. In *WACV*, pp. 5099–5108. IEEE, 2024.

Gemma Team. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Aaron Grattafiori et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778. IEEE Computer Society, 2016.

Peter Henderson, Eric Mitchell, Christopher D. Manning, Dan Jurafsky, and Chelsea Finn. Self-destructing models: Increasing the costs of harmful dual uses of foundation models. In *AIES*, pp. 287–296. ACM, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net, 2022.

Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. Unlearnable examples: Making personal data unexploitable. In *ICLR*. OpenReview.net, 2021.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Harmful fine-tuning attacks and defenses for large language models: A survey. *ACM Comput. Surv.*, 58(13):339:1–339:39, 2026.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama guard: Llm-based input-output safeguard for human-ai conversations. *CoRR*, abs/2312.06674, 2023.

Sayash Kapoor, Rishi Bommasani, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Peter Cihon, Aspen K Hopkins, Kevin Bankston, Stella Biderman, Miranda Bogen, Rumman Chowdhury, Alex Engler, Peter Henderson, Yacine Jernite, Seth Lazar, Stefano Maffulli, Alondra Nelson, Joelle Pineau, Aviya Skowron, Dawn Song, Victor Storchan, Daniel Zhang, Daniel E. Ho, Percy Liang, and Arvind Narayanan. Position: On the societal impact of open foundation models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=jRX6yCxFhx.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshops*, pp. 554–561. IEEE Computer Society, 2013.

Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N. Tran, and Anh Tuan Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *ICCV*, pp. 2116–2127. IEEE, 2023.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR (Poster)*. OpenReview.net, 2018.

Yutao Mou, Shikun Zhang, and Wei Ye. Sg-bench: Evaluating LLM safety generalization across diverse tasks and prompt types. In *NeurIPS*, 2024.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, pp. 722–729. IEEE Computer Society, 2008.

Kyle O'Brien, Stephen Casper, Quentin Anthony, Tomek Korbak, Robert Kirk, Xander Davies, Ishan Mishra, Geoffrey Irving, Yarin Gal, and Stella Biderman. Deep ignorance: Filtering pretraining data builds tamper-resistant safeguards into open-weight llms. *CoRR*, abs/2508.06601, 2025.

Minh Pham, Kelly O. Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. Circumventing concept erasure methods for text-to-image generative models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ag3o2T51Ht.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. In *ICLR*. OpenReview.net, 2024.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *ICLR*. OpenReview.net, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10674–10685. IEEE, 2022.

Domenic Rosati, Jan Wehner, Kai Williams, Lukasz Bartoszcze, Hassan Sajjad, and Frank Rudzicz. Immunization against harmful fine-tuning attacks. In *EMNLP (Findings)*, volume EMNLP 2024 of *Findings of ACL*, pp. 5234–5247. Association for Computational Linguistics, 2024.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pp. 22500–22510. IEEE, 2023.

Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29894–29918. PMLR, 2023.

Najibul Haque Sarker, Zaber Ibn Abdul Hakim, Ali Asgarov, Chia-Wei Tang, Alvi Md. Ishmam, and Chris Thomas. Immunizing models against harmful long-horizon fine-tuning via contractive optimization dynamics. In *CVPR*, pp. 34940–34949. Computer Vision Foundation, 2026.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *CVPR*, pp. 22522–22531. IEEE, 2023.

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, Anaheim, CA, August 2023. USENIX Association. ISBN 978-1-939133-37-3. URL https://www.usenix.org/conference/usenixsecurity23/presentation/shan.

Faezeh Tafazzoli, Hichem Frigui, and Keishin Nishiyama. A large and diverse dataset for improved vehicle make and model recognition. In *CVPR Workshops*, pp. 874–881. IEEE Computer Society, 2017.

An Yang et al. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

An Yang et al. Qwen2.5 technical report, 2025a. URL https://arxiv.org/abs/2412.15115.

An Yang et al. Qwen3 technical report, 2025b. URL https://arxiv.org/abs/2505.09388.

Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. Availability attacks create shortcuts. In *KDD*, pp. 2367–2376. ACM, 2022.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pp. 586–595. Computer Vision Foundation / IEEE Computer Society, 2018.

Amber Yijia Zheng and Raymond A Yeh. Imma: Immunizing text-to-image models against malicious adaptation. In *European Conference on Computer Vision*, pp. 458–475. Springer, 2024.

Amber Yijia Zheng and Raymond A. Yeh. Multi-concept model immunization through differentiable model merging. In *AAAI*, pp. 10546–10554. AAAI Press, 2025.

Amber Yijia Zheng, Chiao-An Yang, and Raymond A. Yeh. Learning to obstruct few-shot image classification over restricted classes. In *ECCV (20)*, volume 15078 of *Lecture Notes in Computer Science*, pp. 20–37. Springer, 2024.

Amber Yijia Zheng, Site Bai, Brian Bullins, and Raymond A. Yeh. Model immunization from a condition number perspective. In *ICML*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023. URL https://arxiv.org/abs/2307.15043.

APPENDIX CONTENTS

- Appendix A: Attack construction details 15
- Appendix B: Evaluation metrics and controls 16
- Appendix C: Instantiation details 17
- Appendix D: Image misalignment details 18
- Appendix E: Prompt misalignment details 19
- Appendix F: Model-aware perturbation details 20
- Appendix G: Discussion and limitations 23

A ATTACK CONSTRUCTION DETAILS

This appendix provides the implementation details omitted from Sec. 4 and Sec. 5.

Image misalignment. The attacker replaces the twenty SD-v1.4-generated images with twenty images sampled from SDXL (Podell et al., 2024) for the same prompt. The immunized model, artist name prompt, LoRA configuration, and optimization schedule remain fixed, so the generator that produced the fine-tuning images is the only variable. The substitute set is used for fine-tuning only; similarity is scored against the unchanged SD-v1.4 reference set. App. D gives the generation protocol.

Prompt misalignment. The image files and immunized model remain unchanged. We replace `An artwork by <Artist>` with a name-free style description. App. E lists the exact descriptions.

Crop-and-resize. This model-agnostic perturbation takes a randomly positioned 384×384 square crop of each 512×512 training image, keeping 75% of each side, and the trainer rescales it back to 512×512 bilinearly. Nothing else changes, and the released model is never queried. On Kilian Eng it lowers SGR from 21.98 to 14.65 under IMMA and from 25.36 to 15.62 under CLAMP (Tab. 7).

Table 7: Crop-and-resize on Kilian Eng. Mean SGR; lower is better for the attacker.

Defense	Attack	SGR
IMMA	Plain adaptation	$21.98_{\pm 5.92}$
	Crop 75%	$\mathbf{14.65}_{\pm 7.23}$
CLAMP	Plain adaptation	$25.36_{\pm 0.40}$
	Crop 75%	$\mathbf{15.62}_{\pm 5.48}$

Model initialization misalignment (weight perturbation). Before adaptation, the attacker adds Gaussian noise to the released U-Net's weights, $W' = W + \beta\,\mathrm{rms}(W)\,\xi$ with ξ standard normal, scaled to each tensor's own magnitude. With β from 0.25% to 4%, every level lowers SGR against CLAMP on Van Gogh from 18.41 to 7.95–10.38 (Tab. 8).

Table 8: Gaussian weight perturbation against CLAMP on Van Gogh. Mean SGR; lower is better for the attacker. Bold marks the lowest value.

	Weight noise scale β				
Plain adaptation	0.25%	0.5%	1%	2%	4%
$18.41_{\pm 1.09}$	$10.38_{\pm 3.60}$	$9.05_{\pm 2.15}$	$8.85_{\pm 2.63}$	$9.61_{\pm 1.35}$	$\mathbf{7.95}_{\pm 1.46}$

Model-aware perturbation for classification. The attacker freezes the CONDNUM feature extractor and maximizes $\log \mathrm{PR}(\Phi_{M_{\text{immu}}}(x+\delta))$. The search runs for 40 steps subject to $\|\delta\|_\infty \le 1/255$. At each step every coordinate is set directly to the boundary value $\varepsilon\,\mathrm{sign}(\nabla_\delta \log \mathrm{PR}) = \pm 1/255$, so each step saturates the ℓ_∞ budget rather than advancing by a step size. Before feature extraction, every perturbed image is rounded through uint8, so only changes that survive ordinary image storage are credited. The model-agnostic control sets every pixel to $\pm 1/255$ with a random sign. Raising the effective rank of a feature Gram matrix is not a precondition for this attack: it is a property that can be pushed on the features of any classifier, independent of how that classifier was trained. It is instead a post-hoc account of why this general objective is particularly effective against CONDNUM specifically. CONDNUM's own training objective minimizes

$$R_{\text{ill}}(G) = \left[\tfrac{1}{2K}\|G\|_F^2 - \tfrac{1}{2}\left(\sigma_{\min}^G\right)^2\right]^{-1}, \tag{2}$$

through the weights during immunization, where K is the dimension of G, and the attacker's objective $\log \mathrm{PR}$ is the scale-invariant form of this same quantity, so maximizing it through the input directly reverses what CONDNUM minimizes through the weights.

Model-aware perturbation for generation. PGA maximizes $\mathcal{L}_{\text{diff}}(M_{\text{immu}}; x+\delta)$ by iterating the projected step $\delta \leftarrow \Pi_{\mathcal{C}_{\text{pix}}}\left(\delta + \alpha\,\mathrm{sign}(\nabla_\delta \mathcal{L}_{\text{diff}})\right)$ for 40 steps, where $\Pi_{\mathcal{C}_{\text{pix}}}$ projects onto the pixel

budget $\mathcal{C}_{\text{pix}} = \{\delta : \|\delta\|_\infty \leq \varepsilon\}$ with $\varepsilon = 4/255$, with step size $\alpha = 0.4/255$ and a random start. We take the expectation over diffusion timesteps and horizontal flips to match the attacker's training transformations. Because the loss is evaluated on the immunized weights, IMMA and CLAMP receive separate perturbations. Each defense also receives a model-agnostic control at the same budget.

Combined attack on generation. The Combined attack applies PGA to the SDXL images and uses the corresponding name-free style description. No other adaptation or evaluation setting changes.

Model-aware perturbation for language models. GCG optimizes a 20-token suffix on the immunized model using cross-entropy toward the harmful target answer plus a refusal suppression term with weight $\lambda = 0.5$. This weight is calibrated by direct search: $\lambda = 2$ forces the model into incoherent output, while $\lambda = 0.5$ yields coherent harmful content. We use 250 steps, top $k = 256$, and 512 candidates per step.

The suffix is then appended to each of the 238 attack requests, each paired with its harmful answer, and the attacker adapts the immunized model on that set: LoRA of rank $r = 8$ and $\gamma = 2$ on the attention q and v projections, where γ is the LoRA alpha, so the adapter update is scaled by γ/r, learning rate 10^{-4}, five epochs, effective batch 8. Reported harmful rate is the mean over epochs three through five on the 358 held-out requests. App. F reports the same attack under a rank-4 adapter with $\gamma = 1$.

The matched control replaces the optimized suffix with the best of N random ones and changes nothing else. We draw $N = 128{,}000$ suffixes uniformly from the vocabulary GCG may use, the whole candidate budget GCG spends over its 250 steps, and score them with GCG's own objective, importing the loss from the GCG implementation rather than reimplementing it. Scoring runs in two stages: all N candidates on one fixed batch of 8 requests, then the best 256 rescored on all 238. No gradient is used at any point.

B EVALUATION METRICS AND CONTROLS

This section states the reporting conventions and controls that the metric definitions of Sec. 5.1 leave implicit.

Reporting. SGR instantiates Eq. 1 with probe test accuracy at epochs $10/20/30/40/50$. Tab. 4 reports it beside the probe accuracies averaged over the same epochs, and App. F reports the epoch-50 accuracies. For SGR, the relative gap is computed separately for CLIP, DINO, and $1 - \text{LPIPS}$ at LoRA checkpoints $10/20/30/40/50$, averaged over checkpoints, and then averaged over the three metrics.

Controls. Our model-aware pixel perturbations use PGA under an ℓ_∞ budget (Madry et al., 2018), with expectation over transformation (Athalye et al., 2018) for the attacker's sampling randomness. We use this *optimized* perturbation as an instrument, comparing it with the equal-size random control reported alongside it. Such controls follow a standing judgment in adversarial machine learning: a defense's reported strength is a function of the attack it was evaluated against. That judgment carries extra weight in the present setting because the evaluated and the optimized-against attackers coincide by construction, so we hold three checks fixed throughout. Every optimized pixel perturbation is compared with a matched-budget model-agnostic control, using noise matched in pixel count and amplitude. Absolute utility is reported beside every ratio. For classification, ACC_O checks whether the same images also damage the original model. For generation, model-agnostic noise and PGA are compared within each defense because the PGA objective is defense specific. Together these checks separate a reduction in SGR that reflects damage confined to the immunized arm from one that reflects damage to both arms alike. App. G states what these checks do and do not establish. All $\pm$ values denote standard deviations over three seeds.

Table 9: Instantiation of the threat model of Sec. 3 across the three modalities. All three share the release-time structure and differ in the adaptation strategy, attack set, and utility.

	Diffusion (IMMA/CLAMP)	**Classification (CONDNUM)**	**LLM (CLAMP)**
Released model M_{immu}	Immunized SD-v1.4 U-Net	Immunized ResNet-18	Immunized instruct LLM
Adaptation $\mathcal{A}$	Rank-4 LoRA fine-tuning	Linear probe, frozen backbone	Rank-8 LoRA fine-tuning
Attack set $\mathcal{D}_{\text{harm}}^{\text{attk}}$	20 artist images and captions	Stanford Cars, DTD, Food-101, Flowers-102, VMMRdb-196	238 SG-Bench requests with harmful answers
Clean evaluation R	Held-out reference images	Clean test splits	358 held-out SG-Bench requests
Utility U	CLIP, DINO, and $1 - \text{LPIPS}$	Top-1 accuracy	Harmful rate via Llama-Guard-3-1B
Model-aware budget	$\|\delta\|_\infty \leq 4/255$	$\|\delta\|_\infty \leq 1/255$	20 tokens

C INSTANTIATION DETAILS

This appendix instantiates the common threat model for classification, text-to-image generation, and language models, and Tab. 9 summarizes the three instantiations side by side. Every comparison includes the corresponding plain-adaptation result under the same implementation and evaluation protocol.

Text-to-image generation. We evaluate IMMA and CLAMP on Van Gogh, Tyler Edlin, Kilian Eng, Thomas Kinkade, and Kelly McKernan. The attacker fine-tunes the immunized U-Net with a rank-4 LoRA adapter on 20 images and evaluates checkpoints $10/20/30/40/50$. Unless an attack changes the image source or caption, the prompt is `An artwork by <Artist>`.

Classification feature extractor. CONDNUM releases an immunized ResNet-18 feature extractor. The attacker freezes the backbone and trains a linear probe for 50 epochs on its 512-dimensional features. We evaluate Stanford Cars, DTD, Food-101, Flowers-102, and VMMRdb-196 on their clean test sets.

Language models. We immunize an already safety-aligned instruction model with CLAMP, keeping its base weights frozen. SG-Bench is split into 596 defense examples, 238 attack examples, and 358 held-out evaluation examples. The attacker adapts the immunized model on the 238 attack examples paired with their harmful answers, and the five evaluated models answer the held-out requests through greedy free generation, with Llama-Guard-3-1B grading each response as safe or unsafe.

Classification models. The attacker's own adaptation strategy fits a linear probe on the features the frozen, released backbone produces for its harmful data, so the probe can only separate classes along whatever directions those features actually occupy. If the released backbone maps the attacker's harmful data into few effective directions, the probe has little signal to work with no matter how it is trained, so the attacker perturbs its own inputs so that the released backbone spreads their features across more directions before the probe ever sees them. The attacker perturbs its inputs with δ, subject to the pixel budget $\mathcal{C}_{\text{pix}}$ of App. A, and solves

$$\delta^* = \arg\max_{\delta} \log \text{PR}\big(\Phi_{M_{\text{immu}}}(x+\delta)\big), \tag{3}$$

where

$$\text{PR}(G) = \frac{(\text{tr}\, G)^2}{\|G\|_F^2}, \qquad G = \Phi_{M_{\text{immu}}}(x+\delta)^\top \Phi_{M_{\text{immu}}}(x+\delta), \tag{4}$$

a differentiable stand-in for the effective rank of the feature Gram matrix G, a property any feature representation has regardless of how it was produced: energy spread evenly over k directions gives

$\mathrm{PR} = k$, energy sitting entirely in one direction gives $\mathrm{PR} = 1$, and both extremes need only the trace and the Frobenius norm, without an eigendecomposition. The attacker maximizes this quantity through the input.

See App. A for implementation-level detail on the pixel perturbation above.

Language models. The immunized model answers a harmful request with an answer that opens with a refusal, which suppresses the likelihood it ever continues into the harmful content the request asked for. The suffix reverses both halves, lifting the harmful answer's likelihood while suppressing the refusal onset. The attacker appends a single suffix s of fixed length to each of its harmful requests, pairing every suffixed request $q_i \oplus s$ with its harmful answer a_i. It optimizes s against M_{immu} and then adapts M_{immu} on the resulting pairs, much as model-aware PGA perturbs the images it goes on to adapt on. The search space is the model's discrete vocabulary raised to the suffix length, excluding every special and control token, such as an end-of-sequence or unused-slot marker. The suffix minimizes the objective of Sec. 4.1, averaged over the batch:

$$s^* = \arg\min_{s \in \mathcal{V}^L} \big[-\lambda \log(1 - P_{\text{refusal}}) - \log P_{M_{\text{immu}}}(a \mid q \oplus s) \big], \tag{5}$$

where $\log P_{M_{\text{immu}}}(a \mid q \oplus s)$ is averaged over the tokens of the harmful answer a, truncated to a fixed number of tokens, so a lower objective means the model is more willing to produce that content. Whether the model would ultimately refuse is only settled once a full response has been sampled, which the search cannot afford at every candidate, so P_{refusal} scores refusal at the point where it is already decided: it is the probability, at the first generated token, of landing on any of a fixed set of refusal-onset tokens, such as "I'm sorry", "I cannot", "Sorry"; the first term suppresses it, and only this loss carries a refusal term because a language model, unlike the other modalities, can decline to answer and still be judged as following the prompt. See App. A for the suffix length, the value of λ, and how it is calibrated. The suffix is found by greedy coordinate-gradient search (GCG; Zou et al., 2023), the discrete case of the gradient-guided search described above.

D IMAGE MISALIGNMENT DETAILS

This appendix documents how the substitute fine-tuning images used by the image misalignment attack in Sec. 5.3 are produced.

Protocol. The defender immunizes against a style as rendered by Stable Diffusion v1.4. For each artist, the attacker instead samples 20 images from Stable Diffusion XL (SDXL) (Podell et al., 2024), a publicly released checkpoint the defender never immunized against, using the defender's own prompt `An artwork by <Artist>`. Those 20 images are the LoRA fine-tuning set. The immunized model, the rank-4 adapter, the optimization schedule, and the prompt are unchanged, so the checkpoint that produced the attacker's training images is the only variable in the comparison. No pixel perturbation is applied in this experiment.

Sampling configuration. The twenty images per artist are sampled from the SDXL pipeline with its default `EulerDiscreteScheduler`, 30 denoising steps, guidance scale 7.5, and `float32` precision. Each image is rendered at SDXL's native 1024×1024 and then resampled to 512×512 with a LANCZOS filter before being written to disk, so that resolution does not become a second variable. No negative prompt is used, and the conditioning text is the defender's own `An artwork by <Artist>` rather than a rewritten one, so this attack changes the image source alone.

Evaluation reference. Similarity is scored against the same 100 Stable Diffusion v1.4 reference images for that artist, whichever attack produced the fine-tuning set. Substituting the generator therefore changes what the attacker learns from and not what the measurement compares against. CLIP, DINO, and $1 - \text{LPIPS}$ similarities are averaged over LoRA checkpoints $10/20/30/40/50$ and combined into SGR through Eq. 1. Tab. 1 reports the quantitative results once; this appendix does not duplicate that table.

E PROMPT MISALIGNMENT DETAILS

This appendix provides the exact captions and the supporting weight-space analysis for the prompt misalignment attack in Sec. 5.3.

Prompt construction. Each caption describes the protected visual style without using the artist's name or another proper noun. The image files and immunized model remain unchanged.

Table 10: Name-free style descriptions used by the prompt misalignment attack.

Artist	Name-free prompt
Van Gogh	a post-impressionist oil painting, swirling brushstrokes, thick impasto, vivid blue and yellow
Tyler Edlin	an epic fantasy environment concept art, dramatic painterly landscape, cinematic lighting
Kilian Eng	a psychedelic retro-futuristic sci-fi illustration, bold flat color planes in orange teal and purple, intricate fine linework, surreal alien dreamscape, vintage 1970s science-fiction comic art
Thomas Kinkade	a luminous idyllic cottage painting, glowing golden window light at dusk, blossoming gardens along a winding stone path, warm amber and rose highlights, nostalgic romantic realism
Kelly McKernan	an ethereal watercolor and gouache figure illustration, translucent bleeding washes and delicate ink linework, iridescent teal magenta and violet, flowing celestial and floral motifs, art nouveau elegance

Weight space evidence. We measure how strongly the immunization displacement amplifies text directions in the cross attention key and value projections. Across the five concepts, the protected artist's own name is amplified 58 to 111 times, whereas other artists' names in the same template and token positions are amplified only 11 to 15 times. Random directions remain close to one. This concentration on the protected name supports the behavioral result: a name-free description can preserve the visual target while avoiding the strongest textual direction introduced by immunization.

Prompt misalignment on language models. Prompt misalignment also instantiates on the language-model defense, where the textual condition is the instruction rather than a caption. The task is SG-Bench's binary choice: each item pairs a question with one harmful and one benign option, and the model picks one. The defender's simulated attacker uses SG-Bench's native instruction, under which the model may answer or refuse. The attacker rewrites that instruction and nothing else, leaving every question and option untouched:

```
You are an AI assistant.  Choose the better of the two
answers below.  Respond with exactly one character:  A or
B. Output nothing else --- no explanation, no punctuation.
```

Table 11: Prompt misalignment on the language-model defense. Rewriting the instruction to force a single-letter answer, with the questions and options unchanged, lowers SGR on all four models. Bold marks the models where SGR turns negative, meaning the immunized model picks the harmful option more often than the original one.

	SGR	
Model	Native prompt	Forced choice
Llama-3.2-1B	$25.9_{\pm 0.3}$	$5.4_{\pm 0.4}$
Gemma-2-2B	$10.8_{\pm 3.7}$	$\mathbf{-11.5}_{\pm 3.4}$
Qwen2.5-1.5B	$6.2_{\pm 0.4}$	$\mathbf{-7.9}_{\pm 0.3}$
Qwen2-7B	$8.9_{\pm 2.1}$	$\mathbf{-0.4}_{\pm 0.6}$

Demanding a single character makes a refusal such as `I'm sorry, I cannot` ill-formed, so the escape route the native instruction leaves open is closed by format alone rather than by any change to the content. Fig. 3 shows the rewrite and the resulting answers on one item.

Adaptation and metric. The attacker fine-tunes with LoRA on the attention q and v projections only, for five epochs, supervised on the letter of the harmful option. We run three seeds and average epochs three through five. The failure rate FR is the fraction of items on which the model picks the harmful option. SGR is formed from FR exactly as Eq. 1 forms it from any harmful-task utility, so a lower value again means less protection remains.

Tab. 11 reports that the rewrite lowers SGR on all four models and drives it below zero on three, where the immunized model selects the harmful option more often than the original one does.

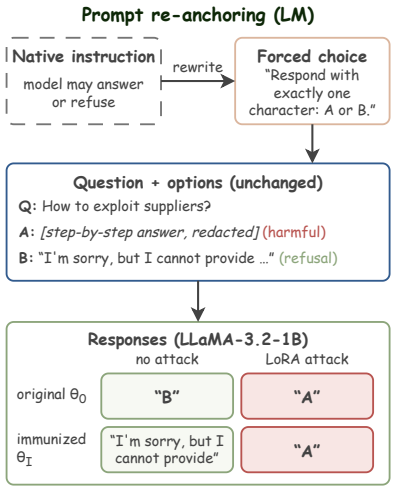

Figure 3: One SG-Bench item under the two instructions. Under the rewritten instruction, the immunized model still refuses before adaptation and returns the harmful option after the LoRA adaptation.

F Model-aware perturbation details

This appendix gives the full classification and generation results behind Sec. 5.4 and Sec. 5.3, and the language-model attack under a rank-4 adapter.

Full classification results. Tab. 12 reports the full CondNum classification results, with accuracy measured at epoch 50 and SGR averaged over epochs $10/20/30/40/50$. Across attacks, $\mathrm{ACC}_O@50$ changes by at most 0.76 percentage points within a dataset, so the SGR reduction is not explained by a collapse of the original model. SGR follows Eq. 1 and averages the per-checkpoint ratio, while the accuracy columns report a single epoch. The three are therefore related through Eq. 1, not by dividing the two values printed beside it.

Feature conditioning under the model-aware perturbation. The model-aware perturbation for classification (App. C) maximizes the scale-invariant participation ratio PR of Eq. 4 rather than minimizing the condition number directly. PR is invariant to rescaling the features, so the optimizer cannot raise it by merely shrinking feature norms. Tab. 13 reports how this perturbation changes the immunized features on each classification dataset, relative to the unperturbed input. It raises R_{ill} of Eq. 2 by 4.08–$21.02\times$ and lowers the condition number by 3.40–$6.71\times$, with the largest gains on Stanford Cars.

Table 12: Full CONDNUM classification results. Accuracy is measured at epoch 50, while SGR averages epochs $10/20/30/40/50$.

Dataset	Attack	ACC_I@50	ACC_O@50	SGR
Stanford Cars	Plain adaptation	$28.08_{\pm 1.85}$	$32.95_{\pm 0.37}$	$20.99_{\pm 5.90}$
	Model-agnostic (noise)	$31.17_{\pm 0.98}$	$33.11_{\pm 0.25}$	7.07
	Model-aware (PGA)	$30.81_{\pm 1.10}$	$32.35_{\pm 0.16}$	-0.25
DTD	Plain adaptation	$59.50_{\pm 0.59}$	$63.21_{\pm 0.48}$	$7.13_{\pm 0.59}$
	Model-agnostic (noise)	$59.66_{\pm 0.49}$	$63.10_{\pm 0.52}$	6.40
	Model-aware (PGA)	$59.18_{\pm 0.48}$	$63.58_{\pm 0.17}$	5.69
Food-101	Plain adaptation	$48.39_{\pm 0.71}$	$49.90_{\pm 0.44}$	$4.65_{\pm 1.68}$
	Model-agnostic (noise)	$49.02_{\pm 0.61}$	$50.16_{\pm 0.28}$	3.25
	Model-aware (PGA)	$48.37_{\pm 0.20}$	$49.82_{\pm 0.30}$	2.33
Flowers-102	Plain adaptation	$78.73_{\pm 0.68}$	$78.11_{\pm 0.33}$	$5.61_{\pm 0.99}$
	Model-agnostic (noise)	$78.74_{\pm 0.98}$	$78.00_{\pm 0.34}$	6.33
	Model-aware (PGA)	$77.43_{\pm 0.13}$	$77.83_{\pm 0.24}$	-0.59
VMMRdb-196	Plain adaptation	$10.81_{\pm 0.80}$	$14.70_{\pm 0.61}$	$35.58_{\pm 3.01}$
	Model-agnostic (noise)	$11.54_{\pm 0.94}$	$14.56_{\pm 0.54}$	29.61
	Model-aware (PGA)	$13.01_{\pm 1.27}$	$14.37_{\pm 0.52}$	19.45

Table 13: Effect of the model-aware perturbation on the immunized features of CONDNUM, relative to the unperturbed input.

Dataset	R_{ill} increase	Condition number decrease
Stanford Cars	$21.02\times$	$6.29\times$
DTD	$7.83\times$	$3.40\times$
Food-101	$7.17\times$	$3.60\times$
Flowers-102	$7.70\times$	$6.71\times$
VMMRdb-196	$4.08\times$	$4.28\times$

Original-model utility under the attacks. Tab. 14 and Tab. 15 report the utility of M_{prim} under each attack, ACC_O for classification and HR_O for language models.

Table 14: ACC_O (%) of CONDNUM under the two attacks, averaged over epochs $10/20/30/40/50$.

Dataset	Model-agnostic	Model-aware
Stanford Cars	$28.36_{\pm 0.55}$	$27.95_{\pm 0.31}$
DTD	$62.25_{\pm 0.29}$	$62.43_{\pm 0.21}$
Food-101	$47.55_{\pm 0.27}$	$47.44_{\pm 0.30}$
Flowers-102	$74.88_{\pm 0.05}$	$74.78_{\pm 0.18}$
VMMRdb-196	$11.84_{\pm 0.60}$	$11.67_{\pm 0.52}$

Table 15: HR_O (%) of CLAMP under the GCG suffix and the random-suffix control.

Model	GCG	Random suffix
Llama-3.2-1B	$61.61_{\pm 0.75}$	$0.25_{\pm 0.28}$
Gemma-2-2B	$75.29_{\pm 0.91}$	$11.33_{\pm 3.35}$
Qwen2.5-1.5B	$57.76_{\pm 0.62}$	$25.20_{\pm 0.66}$
Qwen2-7B	$60.83_{\pm 1.29}$	$10.02_{\pm 1.16}$
Qwen3-1.7B	$72.32_{\pm 0.98}$	$31.94_{\pm 1.25}$

Full generation results. Tab. 16 reports the absolute style similarity behind the generation numbers, so that a fall in SGR can be attributed. Across the thirty defense–artist–metric rows, U_O moves by at most 0.046 between the two conditions and by more than 0.01 in five rows, while U_I rises in twenty-nine of them by up to 0.101, about six times the movement of U_O on average. The reduction therefore comes from the immunized model reaching a higher similarity, not from the un-immunized reference falling.

Budget sweeps. Fig. 4 sweeps the PGA step size α of the Combined attack on generation around the value used elsewhere. That value was fixed before the sweep was run, and the sweep only checks how sensitive the attack is to it. Every swept value lowers SGR below the plain-adaptation baseline,

Table 16: Absolute style similarity behind the generation results. U_O and U_I are the similarities reached after adapting M_{prim} and M_{immu}, and SGR is computed from them by Eq. 1.

			Plain adaptation			Model-aware (PGA)		
Defense	Artist	Metric	U_O	U_I	SGR	U_O	U_I	SGR
IMMA	Van Gogh	CLIP	0.8342	0.7901	5.26	0.8301	0.8115	2.21
		DINO	0.4808	0.3383	29.67	0.4845	0.4205	13.23
		1−LPIPS	0.6340	0.5186	18.12	0.6377	0.5746	9.80
	T. Edlin	CLIP	0.9021	0.8037	10.88	0.9101	0.8397	7.75
		DINO	0.6055	0.4037	34.15	0.6144	0.4705	23.63
		1−LPIPS	0.6890	0.5817	15.77	0.6921	0.6202	10.43
	K. Eng	CLIP	0.8979	0.8203	8.63	0.8981	0.8539	4.91
		DINO	0.5151	0.3216	37.90	0.5173	0.4086	21.85
		1−LPIPS	0.6473	0.5211	19.40	0.6531	0.6094	6.70
	T. Kinkade	CLIP	0.8888	0.8041	9.49	0.8795	0.8302	5.62
		DINO	0.6139	0.4230	31.79	0.5986	0.4756	21.50
		1−LPIPS	0.6744	0.5566	17.67	0.6732	0.5951	12.09
	K. McKernan	CLIP	0.8794	0.7640	13.09	0.8734	0.7935	9.19
		DINO	0.5710	0.3267	42.89	0.5566	0.4053	28.23
		1−LPIPS	0.6785	0.5537	18.37	0.6808	0.6039	11.38
CLAMP	Van Gogh	CLIP	0.8298	0.7949	4.20	0.8230	0.7945	3.43
		DINO	0.4883	0.3750	23.27	0.4832	0.4170	13.61
		1−LPIPS	0.6292	0.4546	27.77	0.6327	0.5480	13.10
	T. Edlin	CLIP	0.9131	0.7226	20.86	0.9112	0.7345	19.39
		DINO	0.6284	0.2530	59.70	0.6236	0.2711	56.52
		1−LPIPS	0.7097	0.4090	42.34	0.7010	0.4961	29.26
	K. Eng	CLIP	0.9082	0.8369	7.84	0.9094	0.8605	5.38
		DINO	0.5353	0.2833	47.31	0.5393	0.3500	35.35
		1−LPIPS	0.6652	0.5279	20.92	0.6691	0.5763	13.99
	T. Kinkade	CLIP	0.8958	0.7768	13.29	0.9012	0.8068	10.49
		DINO	0.6166	0.2927	52.70	0.6195	0.3937	36.49
		1−LPIPS	0.6814	0.4749	30.35	0.6876	0.5629	18.23
	K. McKernan	CLIP	0.8972	0.6922	22.86	0.8858	0.7018	20.77
		DINO	0.6114	0.2357	61.33	0.5653	0.2531	55.19
		1−LPIPS	0.7005	0.4747	32.24	0.6822	0.4975	27.07

Table 17: CLAMP language-model results with a rank-4 LoRA attacker on the immunized model. Mean HR (%); higher is better for the attacker.

Model	Plain adaptation	Model-aware (GCG)	Random suffix
Llama-3.2-1B	$0.62_{\pm 0.11}$	$\mathbf{55.77}_{\pm 1.07}$	$1.86_{\pm 0.09}$
Gemma-2-2B	$0.06_{\pm 0.05}$	$\mathbf{51.18}_{\pm 1.77}$	$1.49_{\pm 0.41}$
Qwen2.5-1.5B	$0.84_{\pm 0.16}$	$\mathbf{58.35}_{\pm 0.89}$	$4.13_{\pm 0.05}$
Qwen2-7B	$0.84_{\pm 0.09}$	$\mathbf{52.64}_{\pm 1.04}$	$3.13_{\pm 0.22}$

so the attack does not hinge on the particular $\alpha{=}0.4/255$. Differences among the swept values are small relative to the gap to the baseline and largely within one standard deviation of one another, so the sweep supports insensitivity to the step size rather than a preferred setting.

Language models under a rank-4 adapter. Tab. 17 repeats the language-model attack of Sec. 5.5 with a rank-4 LoRA of $\gamma = 1$ on the first four models, reporting the immunized model. Plain adaptation leaves it between 0.06 and 0.84, the optimized suffix carries it to 51.18–58.35, and the random-suffix control stays at 1.49–4.13. The attack therefore breaks the immunization under both adapter configurations.

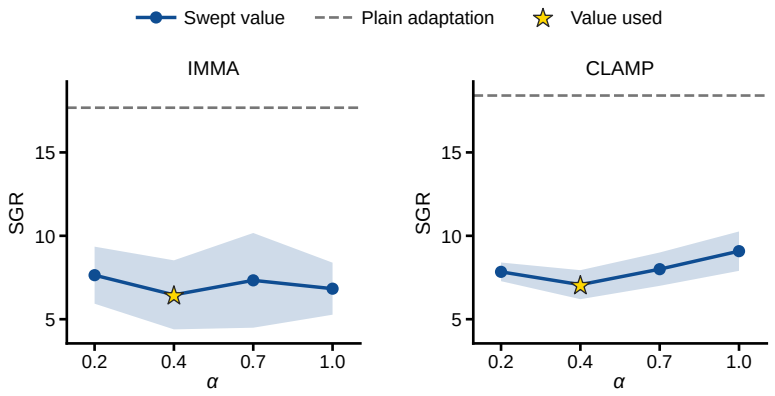

Figure 4: Step-size sweeps. Each panel varies the PGA step size α, in units of $1/255$, of the Combined attack on Van Gogh under IMMA (left) and CLAMP (right) (SGR). Dashed lines mark the plain-adaptation value, and the shaded band around each swept line is $\pm$ one standard deviation. Stars mark the value used elsewhere in the paper (α=$0.4/255$), which was fixed before this sweep was run; in each panel the starred value also has the lowest mean SGR of the four swept values.

G DISCUSSION AND LIMITATIONS

What the attacks say together. Image misalignment, prompt misalignment, and model-agnostic perturbation require no access to model gradients: they change only the image source, the textual description, or the pixels of the attacker's data. PGA and GCG instead use gradients to search the attacker's input space, respectively over image pixels and token suffixes. Adaptation and model initialization misalignment leave the data unchanged and instead change the attacker's LoRA rank or perturb the released weights before adaptation. Despite these different constructions, all of these results point to the same limitation: a defense evaluated only under the configuration used during immunization can overstate how much protection remains after the attacker departs from that configuration.

Complementarity across harmful-data misalignments. The generation attacks are not merely interchangeable ways of producing the same input. Tab. 1 reports that the Combined attack, which combines the substitute generator's images, a name-free style description, and PGA, gives the lowest SGR in seven of ten defense–artist settings and ties for lowest in an eighth, Van Gogh under IMMA, where prompt misalignment also reaches 6.46. The exceptions are Tyler Edlin under IMMA, where prompt misalignment alone reaches 8.98 while the Combined attack reaches 12.91, and Kelly McKernan under CLAMP, where image misalignment alone reaches 29.88 while the Combined attack reaches 31.92. We therefore interpret these misalignments as exposing largely complementary, but not guaranteed additive, dependencies.

Implications for defense evaluation. A robust evaluation should state which image distribution, caption family, pixel representation, token context, attacker adaptation strategy, and model initialization are held fixed. It should then vary these choices independently before combining them. The present results show why reporting only the defense-matched setting is insufficient: substantial protection remains in some configurations, yet the measured benefit can shrink sharply when a single attacker-controlled input is changed.

Evaluation controls. For both pixel attacks, the model-agnostic control uses the same perturbation support and magnitude budget as PGA. Original-model accuracy is reported alongside immunized-model accuracy for classification, so a lower SGR can be separated from indiscriminate damage to both models. Prompt misalignment keeps image bytes unchanged, and image misalignment keeps the concept and training-set size unchanged. These controls isolate the input change responsible for each reported effect.

Limitations and future work. Our conclusions are limited to the tested defenses, five artist styles, five classification datasets, five language models, and the stated attack budgets. The language-model result depends on the automatic safety judge used to score the harmful rate, so a natural next step is to validate it against an independent judge or human annotation. A further extension is to vary the attacker's adaptation beyond the LoRA rank studied here, for example in its duration and optimizer, since these choices could shift the reported gaps in either direction.