DO NOT TRAIN ON MY DATA: UNLEARNABLE DATA VIA CONTRASTIVE REVERSE DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose COREDISTILL (***Co**ntrastive **Re**verse **Distill**ation*), a novel and efficient approach for making image data *unlearnable* from unauthorized generative-model training. Unlike existing methods that iteratively *optimize* each image toward an unlearnability objective, we introduce a sampling-based approach via contrastive decoding that directly *generates* unlearnable samples in a one-shot decoding process. We provide theoretical justification that this sampling procedure produces unlearnable samples, and empirically show that it substantially reduces processing time per image while maintaining strong unlearnability compared with optimization-based baselines. We further observe that the resulting protection transfers across different generative-model variants, providing empirical evidence of cross-model unlearnability. These properties make COREDISTILL practical and scalable for real-world data protection involving diverse generative models and large-scale datasets. This paper contains potentially disturbing images.

1 INTRODUCTION

Recent image generative models have achieved remarkable improvements in visual fidelity and can generate aesthetically compelling images (Wang et al., 2026; Podell et al., 2024; Esser et al., 2024). However, these models are often trained on large-scale datasets collected from the Internet, which may contain copyrighted creative works and other privacy-sensitive content without explicit authorization from the data owners. As model capabilities improve, generative models can also reproduce similar visual characteristics and specific content learned from such training data (Dubiński et al., 2025; Somepalli et al., 2023a; Carlini et al., 2023; Somepalli et al., 2023b; Zou et al., 2025; Zhang et al., 2024). These practices have raised substantial concerns regarding copyright, privacy, and data ownership, and have already led a number of rights holders to pursue legal action against generative-AI developers (Setty, 2023; Brittain, 2024; Chmielewski, 2025; Stempel, 2025).

However, *reactive* legal remedies face practical limitations due to legal uncertainty such as fair use (Milmo, 2025), difficulties in establishing unauthorized use (Jahner, 2026), and the substantial legal resources required. This motivates a *proactive* approach: rather than relying exclusively on remedies after unauthorized training has occurred, data owners should also be able to protect their content *before* it is used for model training. Data cloaking (Shan et al., 2020; 2023) provides a technical basis for such protection by adding visually imperceptible or minimally perceptible perturbations that alter how machine-learning models interpret or learn from an image. For generative-model protection, the goal is therefore to preserve the image's utility to humans while making it difficult for a model to learn from and subsequently reproduce the protected content.

Despite substantial progress, existing cloaking methods still face two practical limitations: *cross-model protection* and *cloaking efficiency*. Most prior methods are developed and evaluated primarily on diffusion-based models(Liang et al., 2023; Van Le et al., 2023; Salman et al., 2023), whereas real-world generative systems may use different architectures and training objectives and are often closed-source (OpenAI, 2024; Google DeepMind, 2026), making the eventual downstream model unknown to data owners. Furthermore, existing methods typically rely on iterative gradient-based optimization over pixels or latent representations (Salman et al., 2023; Van Le et al., 2023; Liang et al., 2023; Zheng et al., 2025), requiring repeated forward and backward passes for each image. This makes them difficult to scale to large real-world datasets, such as tens of thousands of frames from a full-length movie.

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.

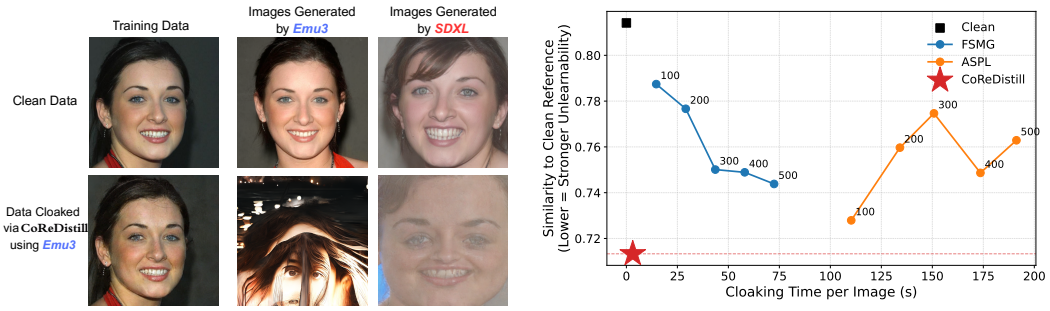

Figure 2: Effectiveness and efficiency of CoReDistill. **Left:** An image cloaked via CoReDistill using Emu3 remains unlearnable to both Emu3 and SDXL. **Right:** CoReDistill achieves strong unlearnability with substantially lower cloaking time, whereas optimization-based baselines require more iterative updates and longer runtime to strengthen unlearnability.

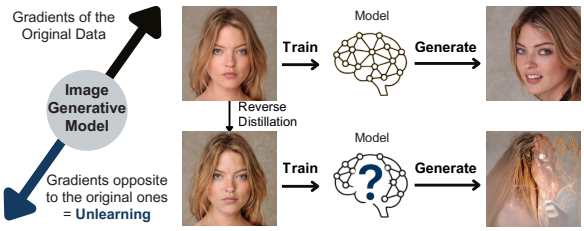

Figure 1: Overview of reverse distillation.

Given these challenges, we propose CoReDistill (***Co**ntrastive **Re**verse **Dis**tillation*), a novel cloaking method designed for efficient large-scale protection with cross-model transfer. Our approach is inspired by *dataset distillation*, which constructs synthetic data that induce learning effects similar to those of the original dataset (Wang et al., 2018; Geng et al., 2023). We reverse this principle by constructing minimally perturbed data whose training gradients oppose those induced by the original datapoints. This opposing *unlearning effect* counteracts learning from the original data, rendering the protected datapoints *unlearnable*, as illustrated in Figure1. Since prior work has shown that distilled learning signals can transfer across models (Zhao et al., 2021; Nguyen et al., 2021; Cazenavette et al., 2022), CoReDistill provides a natural basis for cross-model unlearnability, which we further validate empirically (Figure 2, left).

Directly optimizing the reverse distillation objective is computationally expensive because it requires second-order derivatives. Specifically, constructing an input whose parameter gradient opposes that of the original datapoint requires differentiating the model's parameter gradients with respect to the input. To avoid this cost, we introduce a sampling-based distillation using *contrastive decoding* (Liu et al., 2021; Li et al., 2023; Shi et al., 2024). By contrasting the predictions of a pretrained model and a model fine-tuned on the protected data, we obtain reverse-distilled images in a one-shot decoding/sampling process rather than through iterative image optimization. We show theoretically that images sampled by this contrastive distribution are aligned with the reverse distillation objective (Section 2.3), and empirically demonstrate that this approach achieves substantially lower cloaking time than optimization-based baselines (Figure 2; right).

In this paper, we conduct comprehensive experiments to evaluate CoReDistill from three perspectives. First, we show that our sampling-based formulation achieves comparable or stronger unlearnability than existing baselines and transfers across models with different architectures and training objectives (Section 3.1). Second, we demonstrate that it requires substantially less cloaking time than iterative optimization-based methods (Section 3.2). Third, we show that CoReDistill scales to a real-world copyright-protection scenario involving full-length creative content with tens of thousands of frames (Section 3.3). In summary, our contributions are threefold:

- We propose CoReDistill, a novel cloaking formulation that induces an unlearning effect opposite to the learning effect of the original data, enabling effective cross-model unlearnability.
- We introduce an efficient sampling-based realization of reverse distillation using contrastive decoding, replacing iterative optimization with one-shot decoding and substantially reducing computational cost.
- We demonstrate the practical scalability of CoReDistill by applying it to large-scale copyrighted creative content and validating protection against both in-model and cross-model training.

2 CoReDistill: Cloaking via Contrastive Reverse Distillation

We first review autoregressive image generative models and their next-token prediction objective (Section 2.1). We then formalize reverse distillation and show that directly optimizing its objective requires expensive mixed second-order derivatives (Section 2.2). Finally, we show how contrastive decoding provides an efficient generation-based realization of reverse distillation (Section 2.3).

2.1 Preliminary: Autoregressive Text-to-Image Generative Models

Unlike diffusion- and flow-matching-based text-to-image models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021; Rombach et al., 2022; Liu et al., 2023; Lipman et al., 2023), recent LLM-like image generators adopt autoregressive next-token prediction (Ramesh et al., 2021; Esser et al., 2021; Yu et al., 2022; Chameleon Team, 2024; Wu et al., 2025a). Given an image x, such models first encode it into a sequence of discrete visual tokens using a visual tokenizer, such as a VQ-VAE (van den Oord et al., 2017). Given an image x, its encoder E first maps the image into a sequence of D-dimensional continuous latent vectors:

$$(\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_T) = E(x), \qquad \boldsymbol{h}_t \in \mathbb{R}^D, \tag{1}$$

where T denotes the number of latent representations. Each latent vector $\boldsymbol{h}_t$ is then quantized to its nearest entry in a learned codebook $\mathcal{C} = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{|\mathcal{V}|}\}$:

$$z_t = \arg\min_{v \in \mathcal{V}} \|\boldsymbol{h}_t - \boldsymbol{e}_v\|_2 , \tag{2}$$

where $\mathcal{V}$ denotes the visual-token vocabulary and $\boldsymbol{e}_v$ is the codebook embedding associated with token v. The image is therefore represented by a sequence of discrete visual tokens:

$$\boldsymbol{z} = (z_1, z_2, \ldots, z_T), \qquad z_t \in \mathcal{V}. \tag{3}$$

The autoregressive model $\boldsymbol{\theta}$ is then trained on $\boldsymbol{z}$ using a next-token prediction objective:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z}) = -\log p_{\boldsymbol{\theta}}(\boldsymbol{z}) = -\sum_{t=1}^{T} \log p_{\boldsymbol{\theta}} \left(z_t \mid \boldsymbol{z}_{<t} \right). \tag{4}$$

2.2 Reverse Distillation

Suppose we have a generative model parameterized by $\boldsymbol{\theta}$, a clean image x, and a training objective $\mathcal{L}(\boldsymbol{\theta}, x)$. If the model is trained on x for one step using SGD, its parameters are updated as

$$\boldsymbol{\theta}^* = \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, x), \tag{5}$$

where η denotes the learning rate. Thus, the learning effect induced by x can be characterized by the parameter update $-\eta \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, x)$.

In data cloaking (Shan et al., 2020; 2023), the model is instead trained on a perturbed image $x' = x + \delta$ with $\|\delta\|_p \leq \epsilon$, where δ is an imperceptible perturbation constrained by a budget ϵ. Our key intuition is to choose δ to minimize the dot product between the gradient induced by the cloaked image and the gradient of the clean image:

$$\delta^* = \arg\max_{\|\delta\|_p \leq \epsilon} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, x + \delta)^\top (-\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, x)), \tag{6}$$

Under this condition, training on the cloaked image produces a parameter update opposite to that induced by the clean image. Consequently, the update increases rather than decreases the loss on the corresponding clean image, thereby exerting an *unlearning* effect.

Let $\boldsymbol{\theta}_{\text{cloak}} = \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, x + \delta)$ denote the parameters after one SGD update on the cloaked image. Under a first-order approximation, the clean-image loss change is $\mathcal{L}(\boldsymbol{\theta}_{\text{cloak}}, x) - \mathcal{L}(\boldsymbol{\theta}, x) \approx \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, x)^\top (\boldsymbol{\theta}_{\text{cloak}} - \boldsymbol{\theta}) = -\eta \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, x)^\top \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, x + \delta)$. Thus, maximizing Equation 6 seeks to increase the clean-image loss, thereby discouraging learning on the clean image.

This approach reverses the principle of gradient matching in data distillation (Zhao et al., 2021; Zhao & Bilen, 2021): while gradient matching seeks data that induce a learning effect similar to that of the original data, we seek data that induce the opposite effect. We call this approach *reverse distillation*. Gradient matching is typically optimized via gradient descent, but applying such gradient-based optimization to Equation 6 requires mixed second-order derivatives $\nabla_\delta \nabla_{\boldsymbol{\theta}} \mathcal{L}$, which are computationally expensive for large-scale generative models.

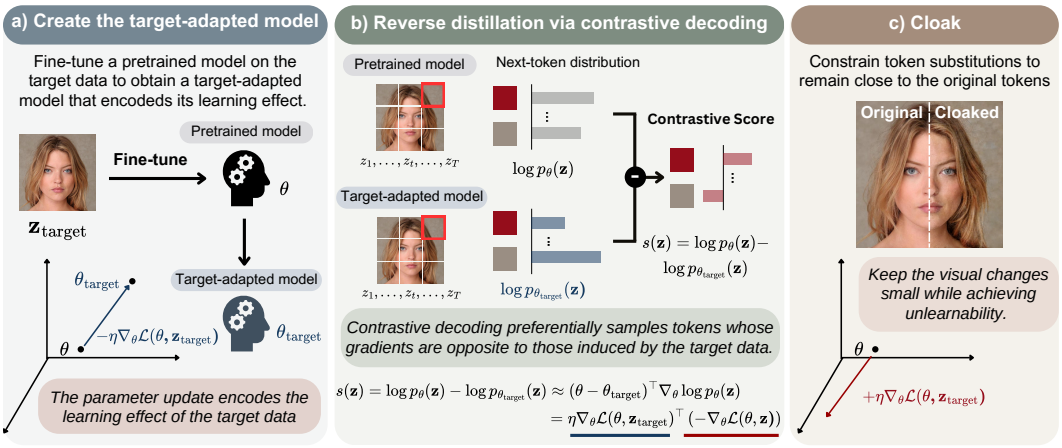

Figure 3: CoReDistill: Reverse Distillation via contrastive decoding.

2.3 CoReDistill: Reverse Distillation via Contrastive Decoding

We now show how contrastive decoding enables the generation-based realization of reverse distillation. Contrastive decoding is a decoding-time technique commonly used to steer the generation of autoregressive models (Liu et al., 2021; Li et al., 2023; Shi et al., 2024). Let $\boldsymbol{\theta}_{\text{target}}$ denote a model fine-tuned on protection target data such as copyrighted data, which we refer to as *target* data, and let $\boldsymbol{\theta}$ denote the pre-trained model. Contrastive decoding assigns a score to a candidate sequence $\boldsymbol{z}$ according to the difference between its log probabilities under the pre-trained and target models:

$$s(\boldsymbol{z}) = \log p_{\boldsymbol{\theta}}(\boldsymbol{z}) - \log p_{\boldsymbol{\theta}_{\text{target}}}(\boldsymbol{z}), \tag{7}$$

$$\boldsymbol{z}^* = \arg\max_{\boldsymbol{z}} s(\boldsymbol{z}). \tag{8}$$

Thus, sequences favored by the target-adapted model are penalized, steering generation away from characteristics learned from the target data at *inference time*.

Interestingly, Lu et al. (2025) showed that data generated through contrastive decoding can also be used at *training time* to unlearn the protection target. Specifically, Lu et al. (2025) found that applying a first-order Taylor expansion of $\log p_{\boldsymbol{\theta}_{\text{target}}}(\boldsymbol{z})$ around $\boldsymbol{\theta}$ in Equation 7 gives

$$s(\boldsymbol{z}) \approx (\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{target}})^\top \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\boldsymbol{z}). \tag{9}$$

Suppose $\boldsymbol{\theta}_{\text{target}}$ is obtained by one SGD update on target data $\boldsymbol{z}_{\text{target}}$: $\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{target}} = \eta\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z}_{\text{target}})$. Since the autoregressive training objective satisfies $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z}) = -\log p_{\boldsymbol{\theta}}(\boldsymbol{z})$, Equation 9 becomes

$$s(\boldsymbol{z}) \approx \eta\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z}_{\text{target}})^\top \left(-\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z})\right). \tag{10}$$

Therefore, maximizing the contrastive score maximizes the dot product between the gradient induced by $\boldsymbol{z}$ and the negative gradient of the target data, which is precisely the reverse distillation objective in Equation 6. Contrastive decoding thus generates reverse-distilled data through a one-shot decoding process, avoiding direct optimization through the expensive second-order derivatives.

In a real-world setting, to preserve the visual appearance of the original image, we additionally constrain token substitutions to remain close to the original visual-token representations. Specifically, we construct the cloaked sample in the following three steps.

Creating the target-adapted model. First, we fine-tune a pre-trained model $\boldsymbol{\theta}$ on the target data we aim to protect, obtaining a target-adapted model $\boldsymbol{\theta}_{\text{target}}$. Under the one-step approximation in Equation 5, the parameter difference between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{\text{target}}$ encodes the learning effect of the protected data, as illustrated in Figure 3 (a).

Reverse distillation via contrastive decoding. As illustrated in Figure 3 (b), given a protected token sequence $\boldsymbol{z} = (z_1, z_2, \ldots, z_T)$, we compute the contrastive score for each candidate visual token $v \in \mathcal{V}$ at position t: $s_t(v) = \log p_{\boldsymbol{\theta}}(v \mid \boldsymbol{z}_{<t}) - \log p_{\boldsymbol{\theta}_{\text{target}}}(v \mid \boldsymbol{z}_{<t})$.

Cloaking. To preserve visual similarity, we restrict candidate substitutions according to their codebook-space distance from the original representation. Let $D_t(v) = \|\boldsymbol{h}_t - \boldsymbol{e}_v\|_2$, where $\boldsymbol{h}_t$ is the original latent representation and $\boldsymbol{e}_v$ is the embedding of candidate token v. We apply an α-mask to the contrastive score:

$$s_t^{(\alpha)}(v) = \begin{cases} s_t(v), & D_t(v) < \alpha \max_{u \in \mathcal{V}} D_t(u), \\ -\infty, & \text{otherwise}, \end{cases} \tag{11}$$

where $\alpha \in [0, 1]$ controls the allowed distance of token substitution. For example, when $\alpha = 0.1$, only candidate tokens whose distance from $\boldsymbol{h}_t$ is less than 10% of the maximum codebook distance at that position remain valid. To limit the number of cloaked visual tokens, we independently draw $m_t \sim \text{Bernoulli}(r)$, where we set $r = 0.5$ by default. If $m_t = 0$, the original token z_t is retained; if $m_t = 1$, the token is substituted using contrastive decoding. The final cloaked token is therefore:

$$z_t^{\text{unlearn}} = \begin{cases} z_t, & m_t = 0, \\ \arg\max_{v \in \mathcal{V}} s_t^{(\alpha)}(v), & m_t = 1. \end{cases} \tag{12}$$

In this way, contrastive decoding favors substitutions with strong unlearnability, while α-masking constrains them to remain visually close to the original representation, as illustrated in Figure 3(c).

3 Experiments

We first evaluate the proposed CoReDistill on two dimensions, effectiveness (Section 3.1) and cloaking efficiency (Section 3.2), compared to previous baselines. After establishing strong cross-model unlearnability and substantially lower per-image cloaking time, we then scale the experiment to large-scale full-length creative contents and further evaluate the proposed method's effectiveness and scalability in a real-world setting (Section 3.3).

3.1 Experiments on Facial Data

3.1.1 Experimental Setups

Datasets. Following previous work (Van Le et al., 2023; Liu et al., 2024b; Zheng et al., 2025), we use CelebA-HQ (Liu et al., 2015; Karras et al., 2018) and the same 50 target identities as Van Le et al. (2023). For each identity, we use four images as reference data to construct the target-adapted model $\boldsymbol{\theta}_{\text{target}}$, and a disjoint set of four unseen images as cloaking targets. This simulates using previously available photos to protect future unseen photos of the same identity.

Training and Cloaking Configurations. We use the autoregressive image generative model Emu3 (Wang et al., 2026) to construct the target-adapted model and generate cloaked images via CoReDistill. Following previous work (Van Le et al., 2023; Zheng et al., 2025), we adopt DreamBooth-style fine-tuning (Ruiz et al., 2023). Each reference image $\boldsymbol{z}_{\text{ref}}$ is paired with $c_{\text{sks}} =$ `"a photo of sks person"`, so that the identifier `"sks"` is associated with the target identity. We fine-tune Emu3 on the reference images to obtain $\boldsymbol{\theta}_{\text{target}}$, and then apply CoReDistill to the disjoint cloaking images using $r = 0.5$ and $\alpha = 0.1$, as described in Equation 12.

For in-model evaluation, we independently fine-tune Emu3 on the clean and Emu3-cloaked images. For cross-model evaluation, we similarly fine-tune SDXL (Podell et al., 2024) on the clean and Emu3-cloaked images. Configurations and hyperparameter tuning are detailed in Section A.1.

Evaluation Metrics. Following previous work (Van Le et al., 2023; Liu et al., 2024b; Zheng et al., 2025), we evaluate each fine-tuned Emu3 and SDXL model using two prompts: the training prompt `"a photo of sks person"` and an unseen test prompt `"a dslr portrait of sks person"`. For each prompt, we generate 32 images and evaluate them using two identity-based metrics. First, we measure the ***Face Detection Failure Rate (FDFR)*** using RetinaFace (Deng et al., 2020), defined as the fraction of generated images in which no face is detected. Second, for images with a detected face, we compute the ***Identity Score Matching (ISM)*** using ArcFace (Deng et al., 2019). Specifically, we extract an ArcFace embedding from each generated face and calculate its cosine similarity to the mean ArcFace embedding of the target identity's clean reference images. Higher FDFR and lower ISM indicate stronger unlearnability.

3.1.2 BASELINE METHODS

FSMG (Anti-DreamBooth). Fully-trained Surrogate Model Guidance (FSMG) (Van Le et al., 2023) first fine-tunes a fixed surrogate (target-adapted) model on reference images of the target identity and then optimizes a perturbation to maximize the target-adapted model's reconstruction loss. Its original objective can be written as $\delta^* = \arg\max_{\|\delta\|_p \le \epsilon} \mathcal{L}(\boldsymbol{\theta}_{\text{target}}, x + \delta)$. Intuitively, FSMG searches for a perturbed example that the target-adapted model assigns a high training loss.

For an autoregressive model, $\mathcal{L}(\boldsymbol{\theta}_{\text{target}}, \boldsymbol{z}) = -\log p_{\boldsymbol{\theta}_{\text{target}}}(\boldsymbol{z})$. We therefore replace the contrastive score in Equation 7 with: $s_t^{\text{FSMG}}(v) = -\log p_{\boldsymbol{\theta}_{\text{target}}}(v \mid \boldsymbol{z}_{<t})$. We then use the same cloaking constraints as COREDISTILL, with $r = 0.5$ and $\alpha = 0.1$ as in Equation 12.

Recall that our method uses $s_t(v) = \log p_{\boldsymbol{\theta}}(v \mid \boldsymbol{z}_{<t}) - \log p_{\boldsymbol{\theta}_{\text{target}}}(v \mid \boldsymbol{z}_{<t})$. FSMG under the discrete setting can be seen as an ablation of our proposed contrastive-decoding objective. Comparing the two therefore isolates the contribution of the base-model term in contrastive decoding.

ASPL (Anti-DreamBooth). Alternating Surrogate and Perturbation Learning (ASPL) (Van Le et al., 2023) further extends FSMG by iteratively updating the model used to guide cloaking.

AdvDM. AdvDM (Liang et al., 2023) constructs adversarial examples by optimizing a perturbation that maximizes the training loss of a fixed pre-trained generative model: $\delta^* = \arg\max_{\|\delta\|_p \le \epsilon} \mathcal{L}(\boldsymbol{\theta}, x + \delta)$. Intuitively, AdvDM searches for a perturbed datapoint that the pre-trained model itself assigns a high training loss. Similarly, we replace the contrastive score in Equation 7 with $s_t^{\text{AdvDM}}(v) = -\log p_{\boldsymbol{\theta}}(v \mid \boldsymbol{z}_{<t})$, and use the same cloaking constraints as COREDISTILL, with $r = 0.5$ and $\alpha = 0.1$ as in Equation 12.

PhotoGuard (Encoder Attack). PhotoGuard (Salman et al., 2023) proposes an encoder attack that perturbs an image such that its latent representation approaches that of a predefined guide image, such as a gray image [1]. Specifically, let x_{guide} denote a guide image and let $(\boldsymbol{h}_1^{\text{guide}}, \ldots, \boldsymbol{h}_T^{\text{guide}}) = E(x_{\text{guide}})$ denote its latent representation. We optimize a pixel-space perturbation by

$$\delta^* = \arg\min_{\|\delta\|_\infty \le \epsilon} \frac{1}{T} \sum_{t=1}^{T} \left\| \boldsymbol{h}_t(x+\delta) - \boldsymbol{h}_t^{\text{guide}} \right\|_2^2, \tag{13}$$

where $(\boldsymbol{h}_1(x+\delta), \ldots, \boldsymbol{h}_T(x+\delta)) = E(x+\delta)$. Baseline implementations are detailed in Section A.1.

Relations between COREDISTILL and the Baselines. FSMG, ASPL, AdvDM, and COREDISTILL are closely related through their training-loss objectives. Recall that FSMG and ASPL maximize $\mathcal{L}(\boldsymbol{\theta}_{\text{target}}, \boldsymbol{z}')$. A first-order expansion around $\boldsymbol{\theta}$ gives

$$\mathcal{L}(\boldsymbol{\theta}_{\text{target}}, \boldsymbol{z}') \approx \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z}') + (\boldsymbol{\theta}_{\text{target}} - \boldsymbol{\theta})^\top \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z}') \tag{14}$$

The first term corresponds to AdvDM's generic loss-maximization objective, while the second captures the target-specific gradient opposition underlying COREDISTILL. Thus, FSMG and ASPL can be viewed as combining generic adversarial and reverse-distillation effects.

3.1.3 RESULTS

As shown in Table 1, COREDISTILL achieves the second-lowest ISM under both prompts in the in-model setting, with larger gains in ISM than in FDFR. This suggests that it primarily suppresses identity-specific characteristics, consistent with the objective decomposition in Equation 14. In the cross-model setting, COREDISTILL is the only protection method that demonstrates unlearnability in the desired direction across all four measurements, achieving the lowest ISM under both prompts. These results highlight the consistent cross-model unlearnability of COREDISTILL.

By contrast, every baseline fails to demonstrate unlearnability on at least one metric, and several measurements are even worse than the clean condition, indicating weak or absent cross-model protection. PhotoGuard is particularly striking: despite providing the strongest in-model protection, it performs worse than clean across all four cross-model measurements.

[1] We omit PhotoGuard+ (Salman et al., 2023) because it targets inference-time image-to-image editing, which falls outside our setting of training-time unlearnability for text-to-image models.

Table 1: In-model and cross-model unlearnability evaluation on facial data. Best and second-best results are shown in **bold** and underlined, respectively.

	`a photo of sks person`		`a dslr portrait of sks person`	
Method	FDFR ↑	ISM ↓	FDFR ↑	ISM ↓
In-model: Generation results of Emu3 fine-tuned on Emu3-cloaked images.				
Clean	0.3256	0.5002	0.3344	0.5072
FSMG	0.4075	0.4456	0.4094	0.4419
ASPL	0.2988	0.4288	0.3088	0.4477
AdvDM	0.3925	0.4345	0.3900	0.3985
PhotoGuard	**0.6381**	**0.1144**	**0.6306**	**0.1053**
RevDistill	0.3500	0.3814	0.3425	0.3891
Cross-model: Generation results of SDXL fine-tuned on Emu3-cloaked images.				
Clean	0.2338	0.5131	0.0269	0.5237
FSMG	0.2194	0.5253	0.0306	0.5504
ASPL	0.2300	0.5081	0.0369	0.5407
AdvDM	**0.2787**	0.5117	0.0350	0.5451
PhotoGuard	0.0756	0.5365	0.0144	0.5488
RevDistill	0.2650	**0.4775**	**0.0519**	**0.5160**



Figure 4: Qualitative results on facial data. The top row shows images used for fine-tuning, either clean or Emu3-cloaked by CoReDistill or a baseline method. The middle and bottom rows show generations from Emu3 (in-model) and SDXL (cross-model) fine-tuned on the corresponding images, respectively.

Figure 4 reveals a clear qualitative contrast between CoReDistill and the baselines. In-model, FSMG, ASPL, and AdvDM disrupt face generation but still preserve recognizable identity-specific features, while PhotoGuard largely collapses generation into unstructured patterns. CoReDistill instead suppresses the protected identity while still producing a plausible generic face, indicating more targeted identity-specific protection. This distinction becomes more pronounced under cross-model evaluation. When transferred to SDXL, the baseline methods largely recover the protected identity, whereas CoReDistill continues to suppress its distinctive facial characteristics.

Overall, CoReDistill provides strong identity-specific protection and more consistent cross-model transfer than the evaluated baselines.

3.2 Cloaking Time Efficiency

In the previous experiments, we adapted optimization-based baselines to Emu3's discrete visual-token space for a controlled comparison. However, these methods are originally designed for continuous image space and require iterative forward and backward passes, whereas CoReDistill uses one-shot decoding. We therefore ask: *how much cloaking time do optimization-based methods require to achieve comparable unlearnability?*

Cloaking Configurations. We use the first identity from the facial-protection experiment in Section 3.1 and measure both average cloaking time per image and downstream unlearnability. For CoReDistill, we directly use the Emu3-cloaked images from Section 3.1. For the optimization-

Table 2: Distributional similarity and generation-quality evaluation on copyrighted creative content.

	Similarity		Generation Quality (CLIP-IQA)				VLM
Model / Finetuning Data	FID ↑	Prec. ↓	Good ↓	Clean ↓	Sharp ↓	Coherent ↓	Fool % ↓
Emu3 / Clean	57.75	0.8440	0.7031	0.8754	0.4763	0.8538	11.4
Emu3 / Emu3-Cloaked	163.18	0.2160	0.4836	0.6285	0.2890	0.6966	2.4
SDXL / Clean	86.54	0.6360	0.4903	0.7808	0.2812	0.7661	6.3
SDXL / Emu3-Cloaked	90.22	0.5240	0.3882	0.7439	0.1905	0.6667	3.5

based baselines, we implement FSMG and ASPL in their original continuous pixel-space formulations using SDXL and vary the total number of PGD updates over $\{100, 200, 300, 400, 500\}$.

The CoReDistill cloaks have a mean absolute pixel-channel difference of $6.66/255$ from the clean images. For FSMG, the differences at 100, 200, 300, 400, and 500 PGD steps are $9.6/255$, $9.7/255$, $9.7/255$, $9.8/255$, and $9.8/255$, respectively. For ASPL, the corresponding differences are $9.6/255$, $9.8/255$, $9.8/255$, $9.8/255$, and $9.9/255$. Thus, the optimization-based baselines introduce larger average pixel changes than CoReDistill across all evaluated budgets.

Evaluation. To evaluate all cloaks on a model distinct from those used to construct them, we use FLUX.1-dev (Black Forest Labs et al., 2025). We independently fine-tune FLUX.1-dev on clean images, CoReDistill-cloaked images, and FSMG- or ASPL-cloaked images under each PGD budget. For each fine-tuned model, we generate 128 images and evaluate unlearnability using ISM.[2] Additional implementation details are provided in Section A.2.

Results. Figure 2 (right) compares downstream unlearnability with cloaking time per image. CoReDistill achieves strong unlearnability in approximately 3 seconds per image, whereas FSMG and ASPL require progressively more PGD updates and substantially longer runtime. Even with 500 updates, neither baseline matches the ISM achieved by CoReDistill.

3.3 Experiments on Copyrighted Creative Content

The preceding experiments demonstrate strong identity-specific protection, cross-model transfer, and efficient cloaking. We next evaluate whether CoReDistill scales to large real-world creative content with commercially valuable domain-specific visual characteristics.

Datasets and Models. We extract 41,600 frames from Disney's *Frozen* (Buck & Lee, 2013) and *Frozen 2* (Buck & Lee, 2019) and caption them with Janus-Pro-7B (Wu et al., 2025a), forming 41,600 frame–caption pairs. We use Emu3 for cloaking and in-model evaluation, and SDXL for cross-model evaluation. For evaluation, we randomly sample 1,000 captions and generate one image per caption from models fine-tuned on clean and cloaked data, respectively. Dataset construction and training details are provided in Section A.3.

Evaluation Metrics. We evaluate the generated images from three complementary perspectives. First, we measure **distributional similarity** to the original Frozen frames using FID (Heusel et al., 2017) and Precision (Kynkäänniemi et al., 2019); higher FID and lower Precision indicate stronger protection. Second, we evaluate **generation quality** using CLIP-IQA (Wang et al., 2023), with the adjective pairs `good/bad`, `clean/noisy`, `sharp/blurry`, and `coherent/distorted`; lower scores indicate stronger degradation. Finally, we evaluate how easily generated images can be distinguished from genuine Frozen frames using an external VLM judge. Specifically, we present the VLM[3] with the original Frozen frame and the corresponding generated image, and ask it to identify which image is AI-generated. We define the **VLM Judge Fool Rate** as the fraction of cases in which the judge incorrectly identifies the genuine frame as AI-generated. A lower fool rate indicates that the generated image is easier to distinguish and therefore reflects stronger protection.

[2]FLUX.1-dev generates detectable human faces in all evaluated settings, resulting in an FDFR of zero.

[3]Gemini-3-Flash-Preview. https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview

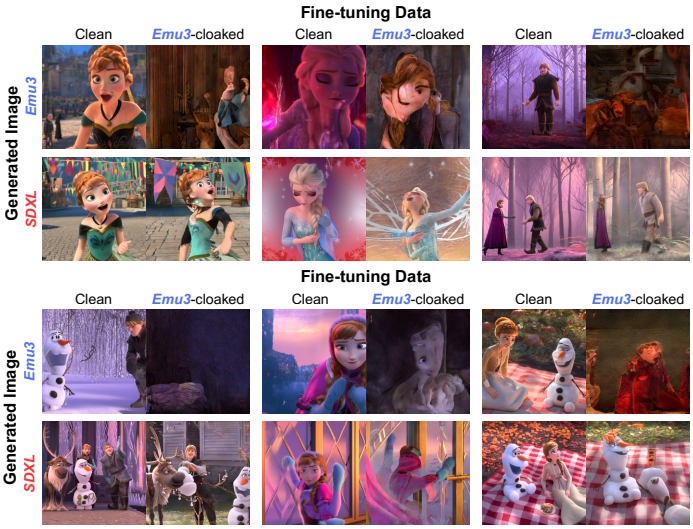

Figure 5: Qualitative generation results on copyrighted creative content. Each block shows generations conditioned on the same caption. Columns compare models fine-tuned on clean and Emu3-cloaked movie frames, while rows compare the in-model setting (Emu3) and cross-model setting (SDXL).

Results. Figure 5 and Table 2 show strong in-model protection. Compared with Emu3 trained on clean frames, training on cloaked frames substantially degrades character identity, facial structure, and scene coherence. Consistently, FID increases markedly, while Precision, all CLIP-IQA scores, and VLM Judge Fool Rate decrease.

Cross-model protection on SDXL is weaker but remains consistent. Although the broad scene semantics are often preserved, generations exhibit degraded character appearance and structural details, with all quantitative metrics moving in the direction of stronger protection. Overall, CoReDistill scales to large-scale copyrighted content while retaining measurable protection across different models.

4 Related Work

Adversarial attacks (Szegedy et al., 2014; Goodfellow et al., 2015; Madry et al., 2018) were originally developed to manipulate model predictions and were later adapted for proactive data protection. Fawkes (Shan et al., 2020) introduced image cloaking against unauthorized face-recognition training, while unlearnable examples (Huang et al., 2021) aim to make released data difficult for neural networks to learn. With the rise of text-to-image models, perturbation-based protection was extended to unauthorized generation and customization. AdvDM (Liang et al., 2023) maximizes the diffusion training loss of a pre-trained model, while Anti-DreamBooth (Van Le et al., 2023) maximizes training loss under fixed or iteratively updated surrogate models. Targeted methods include Glaze (Shan et al., 2023), which shifts artistic representations toward a decoy style; PhotoGuard (Salman et al., 2023), which redirects image representations toward a predefined target image; and ACE (Zheng et al., 2025), which makes models learn a predefined chaotic target pattern. Nightshade (Shan et al., 2024) instead performs prompt-specific poisoning at the concept level.

MetaCloak (Liu et al., 2024b) uses meta-learning and multiple surrogates to improve transferability and robustness; SimAC (Wang et al., 2024a) exploits timestep-dependent properties of diffusion models; MYOPIA (Wu et al., 2025b) improves imperceptibility and robustness; and InMark (Liu et al., 2024a) places protective watermarks on influential regions to resist real-world transformations. Most closely related to our focus on efficiency, Xue et al. (2024) use Score Distillation Sampling (Poole et al., 2023) to avoid expensive backpropagation through the diffusion denoiser, but still require iterative optimization for each image. CoReDistill instead introduces a fundamentally different approach, a generation-based paradigm that directly constructs protected samples through one-shot decoding, shifting data cloaking from *optimization-based* to *generation-based protection*.

5 Conclusion

We proposed CoReDistill, a novel generation-based cloaking method that directly produces unlearnable data via contrastive decoding. Experiments demonstrate substantial reductions in cloaking time over optimization-based methods while providing empirical evidence of cross-model transfer. We hope this work contributes to practical data protection and a more sustainable GenAI ecosystem.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks.

Required disclosures. For synthetic data generation, the frame–caption dataset used in Section 3.3 was captioned with Janus-Pro-7B, as described in Section 3.3 and detailed in Section A.3.

For AI-assisted evaluation, we use Gemini-3-Flash-Preview as a VLM judge to distinguish generated images from genuine movie frames, as described in Section 3.3 and detailed in Section A.3. Its judgments are aggregated into the VLM Judge Fool Rate reported in our experiments. The VLM is used only for evaluation.

We also used ChatGPT (GPT-5.6 Sol [4]) to improve the readability and linguistic quality of the manuscript.

Recommended disclosures. We used ChatGPT (GPT-5.6 Sol) to suggest possible layouts for some figures in the paper. These suggestions were used only as design references; all figures were manually created by the authors using Canva [5].

For software implementation, we first referred to the official open-source implementations of the baseline methods and used generative AI tools to assist with adapting portions of the code to our experimental setting and computing environment [6] [7] [8] . All AI-assisted code was inspected and tested by the authors.

All AI-assisted outputs were reviewed by the authors, who take full responsibility for the final content of the paper, including its text, mathematical claims, experimental methodology, code, results, and figures.

ETHICS STATEMENT

This work aims to help creators and individuals protect their images from being used to train or customize generative models without permission. The proposed method is intended for defensive use: it modifies images so that models have greater difficulty learning from them. Because the technique is related to adversarial perturbation and data poisoning, it could in principle be repurposed in unintended settings. We therefore limit our study to content-protection scenarios.

Our experiments use publicly available datasets and selected copyrighted material only for research evaluation. In particular, frames from *Frozen* and *Frozen2* are used solely to evaluate whether the proposed method can protect real-world creative content from unauthorized learning, with attribution to the original work. We do not claim ownership of any third-party copyrighted material. After completing the study, we will delete the copyrighted experimental data and model checkpoints that can reproduce such content.

We also use publicly available facial images from CelebA-HQ to evaluate whether the proposed protection prevents a model from learning and reproducing identity-specific characteristics. Identity similarity is used only as an evaluation metric; we do not use the facial data for sensitive-attribute inference, profiling, or other biometric analysis.

REPRODUCIBILITY STATEMENT

All experiments were conducted using eight NVIDIA H200 GPUs. For the facial-data experiments (Section 3.1 and Section 3.2), we will release all datasets and source code required to reproduce our results. For the experiments involving copyrighted content (Section 3.3), we are unable to redistribute the experimental data due to copyright restrictions. Nevertheless, we provide detailed descriptions of the experimental setup, implementation, and evaluation procedures in Section 3.3 and Section A.3 to facilitate reproduction of these experiments.

[4] https://openai.com/index/gpt-5-6/

[5] https://www.canva.com/

[6] FSMG and ASPL: `https://github.com/VinAIResearch/Anti-DreamBooth`

[7] AdvDM: `https://github.com/psyker-team/mist`

[8] PhotoGuard: `https://github.com/MadryLab/photoguard`

REFERENCES

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Blake Brittain. Stability AI, midjourney should face artists' copyright case, judge says. *Reuters*, May 2024.

Chris Buck and Jennifer Lee. Frozen. Film, Walt Disney Animation Studios, 2013.

Chris Buck and Jennifer Lee. Frozen II. Film, Walt Disney Animation Studios, 2019.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 2023)*, pp. 5253–5270. USENIX Association, 2023.

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A. Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, pp. 10718–10727, 2022.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Dawn Chmielewski. Disney, universal sue image creator midjourney for copyright infringement. *Reuters*, June 2025.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, pp. 4690–4699, 2019.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)*, pp. 5203–5212, 2020.

Jan Dubiński, Antoni Kowalczuk, Franziska Boenisch, and Adam Dziedzic. CDI: Copyrighted data identification in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025)*, pp. 18674–18684, 2025.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021)*, pp. 12873–12883, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, volume 235, pp. 12606–12633. PMLR, 2024.

Jiahui Geng, Zongxiong Chen, Yuandou Wang, Herbert Woisetschläger, Sonja Schimmler, Ruben Mayer, Zhiming Zhao, and Chunming Rong. A survey on dataset distillation: Approaches, applications and future directions. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI 2023)*, pp. 6610–6618, 2023.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *The Third International Conference on Learning Representations (ICLR 2015)*, 2015.

Google DeepMind. Gemini 3.1 Flash Image model card. Google DeepMind Model Card, 2026.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*, pp. 7514–7528. Association for Computational Linguistics, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS 2017)*, volume 30, pp. 6626–6637, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, volume 33, pp. 6840–6851, 2020.

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 *The Tenth International Conference on Learning Representations (ICLR 2022)*, 2022.

Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. Unlearnable examples: Making personal data unexploitable. In *The Ninth International Conference on Learning Representations (ICLR 2021)*, 2021.

Kyle Jahner. OpenAI must turn over 20 million ChatGPT logs, judge affirms. *Bloomberg Law*, January 2026.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *The Sixth International Conference on Learning Representations (ICLR 2018)*, 2018.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems (NeurIPS 2019)*, volume 32, 2019.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), Volume 1: Long Papers*, pp. 12286–12312. Association for Computational Linguistics, 2023.

Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, volume 202, pp. 20763–20786. PMLR, 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*, 2023.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021), Volume 1: Long Papers*, pp. 6691–6706. Association for Computational Linguistics, 2021.

Hanwen Liu, Zhicheng Sun, and Yadong Mu. Countering personalized text-to-image generation with influence watermarks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*, pp. 12257–12267, 2024a.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*, 2023.

Yixin Liu, Chenrui Fan, Yutong Dai, Xun Chen, Pan Zhou, and Lichao Sun. MetaCloak: Preventing unauthorized subject-driven text-to-image diffusion-based synthesis via meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*, pp. 24219–24228, 2024b.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2015)*, pp. 3730–3738, 2015.

Huimin Lu, Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. UniDetox: Universal detoxification of large language models via dataset distillation. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, 2025.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *The Sixth International Conference on Learning Representations (ICLR 2018)*, 2018.

Dan Milmo. Anthropic did not breach copyright when training AI on books without permission, court rules. *The Guardian*, June 2025.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.

Timothy Nguyen, Zhourong Chen, and Jaehoon Lee. Dataset meta-learning from kernel ridge-regression. In *The Ninth International Conference on Learning Representations (ICLR 2021)*, 2021.

OpenAI. GPT-4o system card. OpenAI System Card, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2023)*, pp. 4195–4205, 2023.

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 *The Twelfth International Conference on Learning Representations (ICLR 2024)*, 2024.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*, 2023.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, volume 139, pp. 8821–8831. PMLR, 2021.

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 2022)*, pp. 10684–10695, 2022.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023)*, pp. 22500–22510, 2023.

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 (ICML 2023)*, volume 202, pp. 29894–29918. PMLR, 2023.

Riddhi Setty. Getty images sues stability AI over art generator IP violations. *Bloomberg Law*, February 2023.

Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y. Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th USENIX Security Symposium (USENIX Security 2020)*, pp. 1589–1604. USENIX Association, 2020.

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 2023)*, pp. 2187–2204. USENIX Association, 2023.

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 *45th IEEE Symposium on Security and Privacy (IEEE S&P 2024)*, pp. 807–825, 2024.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2024), Volume 2: Short Papers*, pp. 783–791. Association for Computational Linguistics, 2024.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, volume 37, pp. 2256–2265. PMLR, 2015.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023)*, pp. 6048–6058, 2023a.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS 2023)*, volume 36, pp. 47783–47803, 2023b.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *The Ninth International Conference on Learning Representations (ICLR 2021)*, 2021.

Jonathan Stempel. Warner bros discovery sues AI photo generator midjourney for stealing superman, scooby-doo. *Reuters*, September 2025.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *The Second International Conference on Learning Representations (ICLR 2014)*, 2014.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS 2017)*, volume 30, 2017.

Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N. Tran, and Anh Tran. Anti-DreamBooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2023)*, pp. 2116–2127, 2023.

Feifei Wang, Zhentao Tan, Tianyi Wei, Yue Wu, and Qidong Huang. SimAC: A simple anti-customization method for protecting face privacy against text-to-image synthesis of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*, pp. 12047–12056, 2024a.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2023)*, volume 37, pp. 2555–2563, 2023.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024), Volume 1: Long Papers*, pp. 9440–9450. Association for Computational Linguistics, 2024b.

Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

Xinlong Wang, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Zhen Li, Yuqi Wang, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Chunlei Men, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Zhongyuan Wang, and Tiejun Huang. Multimodal learning with next-token prediction for large multimodal models. *Nature*, 650:327–333, 2026.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025)*, pp. 12966–12977, 2025a.

Zhihao Wu, Yushi Cheng, Tianyang Sun, Xiaoyu Ji, and Wenyuan Xu. MYOPIA: Protecting face privacy from malicious personalized text-to-image synthesis via unlearnable examples. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2025)*, volume 39, pp. 905–913, 2025b.

Haotian Xue, Chumeng Liang, Xiaoyu Wu, and Yongxin Chen. Toward effective protection against diffusion-based mimicry through score distillation. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, 2024.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022.

Yang Zhang, Tze Tzun Teoh, Wei Hern Lim, and Kenji Kawaguchi. On copyright risks of text-to-image diffusion models. In *The Dark Side of Generative AIs and Beyond Workshop at the European Conference on Computer Vision (ECCV 2024)*, 2024.

Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, volume 139, pp. 12674–12685. PMLR, 2021.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *The Ninth International Conference on Learning Representations (ICLR 2021)*, 2021.

Boyang Zheng, Chumeng Liang, and Xiaoyu Wu. Targeted attack improves protection against unauthorized diffusion customization. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS 2023)*, volume 36, pp. 46595–46623, 2023.

Zihang Zou, Boqing Gong, and Liqiang Wang. Attention to neural plagiarism: Diffusion models can plagiarize your copyrighted images! In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2025)*, pp. 19546–19556, 2025.

A Experimental Details

A.1 Experiments on Facial Data

Hyperparameter Tuning. We use the first identity in the dataset ("identity 5") for training configuration hyperparameter tuning. We first tune the target-adapted Emu3 and select a checkpoint that can reliably learn and reproduce the reference data, which is then used to construct CoReDistill. Separately, we tune the downstream fine-tuning configurations of Emu3 and SDXL on clean, unprotected images to ensure that both models can reliably learn and reproduce the target identity. All selected configurations are then fixed for all the other identities.

Training and Cloaking Configurations. To construct the target-adapted model, each reference image $\boldsymbol{z}_{\text{ref}}$ is paired with $c_{\text{sks}} =$ `"a photo of sks person"`, while a generic person image $\boldsymbol{z}_{\text{generic}}$ is paired with $c_{\text{generic}} =$ `"a photo of person"`. The generic images are generated by the same pre-trained Emu3 model and serve as prior-preservation examples for maintaining the model's general concept of a person. The DreamBooth training objective is

$$\mathcal{L}_{\text{DreamBooth}} = \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z}_{\text{ref}}, c_{\text{sks}}) + \lambda \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z}_{\text{generic}}, c_{\text{generic}}), \tag{15}$$

where λ denotes the prior-preservation weight.

We first fine-tune Emu3 on the reference images using Equation 15 to obtain the target-adapted model $\boldsymbol{\theta}_{\text{target}}$.

To construct the target-adapted Emu3, each optimization step contains four reference images from the target identity and four generic person images, for a total batch size of eight. We train Emu3 for 100 optimization steps using a polynomial learning-rate scheduler with power 1.0, with the learning rate decreasing from 5×10^{-5} to 4×10^{-5}. [9] The prior-preservation weight is set to $\lambda = 1$, following previous work (Van Le et al., 2023). The resulting model is used as the target-adapted model $\boldsymbol{\theta}_{\text{target}}$ for CoReDistill.

We then cloak the second, disjoint set of four images from the same identity using CoReDistill with $r = 0.5$ and $\alpha = 0.1$, as described in Equation 12. These cloaking images are not used to construct the target-adapted model.

For the in-model evaluation, we independently fine-tune two copies of the pre-trained Emu3 model on either the four clean images or their cloaked counterparts. Both models are trained for 100 steps with a constant learning rate of 5×10^{-5} using Equation 15 with $\lambda = 1$. This configuration allows the model trained on clean images to reliably reproduce the target identity, providing a meaningful reference for evaluating unlearnability.

For the cross-model evaluation, we similarly fine-tune SDXL on either the clean or Emu3-cloaked images. We empirically find that SDXL can learn the target identity effectively from only four images without prior preservation, whereas prior preservation degrades identity learning in this setting. We therefore set $\lambda = 0$ for SDXL. We jointly fine-tune the text encoder and U-Net for 100 steps using constant learning rates of 5×10^{-6} and 1×10^{-5}, respectively.

Implementation Details of the Baselines. For FSMG, ASPL, and AdvDM, we apply the same token-position and visual-similarity constraints used by CoReDistill. Specifically, we set $r = 0.5$ and $\alpha = 0.1$ as in Equation 12, while substituting the contrastive score with the corresponding baseline score.

For FSMG, we use the same target-adapted Emu3 employed by CoReDistill as the fixed surrogate model.

For ASPL, following the released implementation of Anti-DreamBooth, the original method uses 50 outer iterations with three surrogate-model updates per iteration, corresponding to 150 surrogate updates in total. This is approximately 15% of the 1,000 optimization steps used to train the clean reference model in the original setting. To preserve a comparable relative update budget in our 100-step setting, we therefore allocate 15 surrogate updates in total. We keep three surrogate-model updates per outer iteration, as in the original implementation, and therefore use five outer iterations.

For PhotoGuard, we optimize the image in continuous pixel space with an ℓ_∞ perturbation budget of $16/255$, a PGD step size of $2/255$, and 200 optimization steps, following the original implementation.

A.1.1 Additional Results.

We additionally report CLIP-IQA (Wang et al., 2023) and BRISQUE (Mittal et al., 2012) results in Table 3 and Table 4. For CLIP-IQA, we evaluate the adjective pairs `good/bad`, `clean/noisy`, `sharp/blurry`, and `coherent/distorted`.

We omit these metrics from the main paper because, in our facial experiments, they do not consistently reflect perceptual generation quality, particularly in the in-model setting. Figure 6 illustrates this issue by showing images generated by Emu3 fine-tuned on clean data and on data cloaked with CoReDistill, together with their corresponding CLIP-IQA and BRISQUE scores. Although the generations from the cloaked-data model are qualitatively far more distorted and often visually incoherent, both CLIP-IQA and BRISQUE can nevertheless assign them better scores. We therefore

[9] During hyperparameter tuning, we empirically found that checkpoint 100 of a 500-step linear-decay schedule initialized at 5×10^{-5} reproduces the reference data well without evident underfitting or overfitting.

Table 3: In-model generation-quality evaluation on facial data. The table reports results of Emu3 fine-tuned on images cloaked using the same Emu3 model. For CLIP-IQA, G/B, C/N, S/B, and C/D denote Good vs. Bad, Clean vs. Noisy, Sharp vs. Blurry, and Coherent vs. Distorted, respectively.

	`a photo of sks person`					`a dslr portrait of sks person`				
Method	G/B ↓	C/N ↓	S/B ↓	C/D ↓	BRI ↑	G/B ↓	C/N ↓	S/B ↓	C/D ↓	BRI ↑
Clean	0.52	0.50	0.48	0.69	34.57	0.53	0.51	0.49	0.70	35.32
FSMG	0.53	0.45	0.51	0.65	36.92	0.53	0.45	0.51	0.66	36.52
ASPL	0.55	0.47	0.59	0.65	36.92	0.54	0.46	0.58	0.65	37.67
AdvDM	0.56	0.47	0.56	0.66	36.45	0.56	0.47	0.56	0.65	37.56
PhotoGuard	0.36	0.39	0.31	0.64	57.23	0.37	0.39	0.32	0.63	56.65
RevDistill	0.59	0.48	0.59	0.62	31.08	0.59	0.47	0.61	0.63	31.28

Table 4: Cross-model generation-quality evaluation on facial data. The table reports results of SDXL fine-tuned on images cloaked using Emu3. For CLIP-IQA, G/B, C/N, S/B, and C/D denote Good vs. Bad, Clean vs. Noisy, Sharp vs. Blurry, and Coherent vs. Distorted, respectively.

	`a photo of sks person`					`a dslr portrait of sks person`				
Method	G/B ↓	C/N ↓	S/B ↓	C/D ↓	BRI ↑	G/B ↓	C/N ↓	S/B ↓	C/D ↓	BRI ↑
Clean	0.41	0.47	0.35	0.83	51.63	0.50	0.59	0.43	0.81	44.36
FSMG	0.38	0.45	0.29	0.77	51.13	0.44	0.54	0.38	0.75	46.57
ASPL	0.39	0.45	0.33	0.78	51.62	0.48	0.55	0.44	0.77	46.05
AdvDM	0.37	0.42	0.28	0.78	52.82	0.46	0.53	0.39	0.75	46.47
PhotoGuard	0.20	0.18	0.18	0.51	35.25	0.21	0.19	0.19	0.49	33.05
RevDistill	0.37	0.44	0.28	0.77	50.98	0.44	0.54	0.36	0.74	45.98

treat these metrics as supplementary rather than primary measures of unlearnability in the facial experiments.

A.2 Experiments on Cloaking Time

Cloaking Configurations. CoReDistill uses the same Emu3 configuration as in Section 3.1, with $r = 0.5$ and $\alpha = 0.1$ following Equation 12.

For FSMG and ASPL, we use their original continuous pixel-space formulations with SDXL and optimize the perturbations using PGD with an ℓ_∞ budget of $16/255$ and a step size of $2/255$. For FSMG, we vary the number of PGD updates over $\{100, 200, 300, 400, 500\}$.

For ASPL, following the original implementation, we use 50 outer iterations and three surrogate-model updates per outer iteration. We vary the number of PGD updates per outer iteration over $\{2, 4, 6, 8, 10\}$, corresponding to total PGD budgets of $\{100, 200, 300, 400, 500\}$, respectively. For each configuration, we record the average cloaking time per image.

Downstream Evaluation Model. We use FLUX.1-dev (Black Forest Labs et al., 2025) as an independent downstream evaluation model. FLUX.1-dev differs from both cloaking models: it uses flow matching rather than next-token prediction or diffusion score matching, and adopts a DiT-based architecture (Peebles & Xie, 2023) rather than the autoregressive transformer of Emu3 or the U-Net architecture of SDXL.

For each cloaking condition, we independently fine-tune a LoRA adapter (Hu et al., 2022) on FLUX.1-dev. All adapters are trained for 500 steps with a batch size of 4 using the Prodigy optimizer and a constant learning rate of 1.0. For each fine-tuned adapter, we generate 128 images using the prompt "`a photo of sks person`".

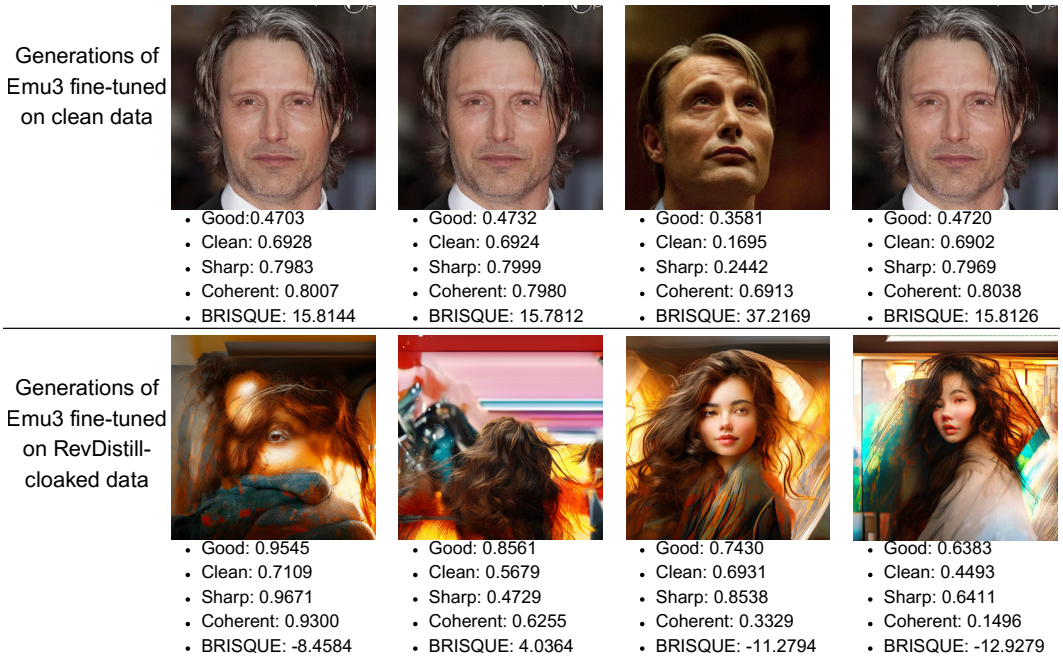

Figure 6: Examples illustrating the limitations of CLIP-IQA and BRISQUE in our facial experiments. Images are generated by Emu3 fine-tuned on clean data (**top**) or data cloaked with CoReDistill (**bottom**). Good, Clean, Sharp, and Coh. denote the four CLIP-IQA dimensions, for which higher values indicate better predicted quality; lower BRISQUE indicates better predicted quality. Despite substantially stronger visual distortions in the cloaked-data generations, these metrics can assign them comparable or even better scores.

A.3 Experiments on Copyrighted Creative Content

Datasets and Models. Each frame is captioned using Janus-Pro-7B (Wu et al., 2025a) with the following prompt: `"Look at the image and write a caption in 3--5 sentences (at most 80--100 words). Describe the scene, any characters, their actions, expressions, clothing, the environment, colors, lighting, and mood."`

Training and Cloaking Configurations. Unlike the facial-data setting, the Frozen dataset is sufficiently large and diverse to support direct fine-tuning without DreamBooth-style prior preservation. We therefore fine-tune Emu3 directly on the clean frame–caption pairs for three epochs with a batch size of 8 and a linear learning rate of 5×10^{-5}, obtaining the target-adapted Emu3 used for CoReDistill.

The protection setting also differs from the facial scenario. For facial data, previously available images are used to construct the target-adapted model, which is then used to cloak future unseen photos. For copyrighted creative content, however, the content itself can be protected before release. We therefore train the target-adapted model on the Frozen dataset and subsequently cloak the same dataset using CoReDistill with $r = 0.5$ and $\alpha = 0.1$, following Equation 12.

To evaluate in-model unlearnability, we independently fine-tune Emu3 on the clean and cloaked Frozen datasets using the same training configuration as above. To evaluate cross-model unlearnability, we independently fine-tune SDXL on the clean and Emu3-cloaked datasets for three epochs with a batch size of 8 and a linear learning rate of 5×10^{-6}. This time, since there is no special string like `"sks"`, we train only the U-net of SDXL.

VLM Judge Fool Rate. For each evaluation example, we present the judge with the original Frozen frame and the image generated from its corresponding caption. The judge receives the following instruction:

```
You are shown two images associated with this caption:
"{caption}"
One image is the original Disney's Frozen movie still that
was captioned by an image-captioning model.
The other image was generated by an AI image generation
model fine-tuned on Frozen movie stills, using the same
caption as its prompt.
Your task is to identify which image is AI-generated.
Answer with exactly one character:
0 if the FIRST image is AI-generated.
1 if the SECOND image is AI-generated.
Do not explain.  Do not output any other text.
```

To eliminate positional bias (Wang et al., 2024b; Zheng et al., 2023), we randomly shuffle the order of the original and generated images independently for each query.

Captions Used in Qualitative results. We use the following captions to generate images used in Figure 5 (from left to right, from upper to lower):

```
"A scene from Disney's Frozen is depicted, showing an
animated character with bright red hair styled in braids,
wearing a green and black dress with intricate patterns
and a pendant necklace.  The character looks surprised or
excited, with wide eyes and an open mouth.  The background
features a quaint village setting with festive decorations
like colorful flags and lanterns, adding to the cheerful
atmosphere.  The environment is warm and vibrant, with soft
lighting enhancing the scene's magical and lively mood."
"A scene from Disney's Frozen, the image shows the character
Elsa with her eyes closed, gently touching a snowflake with
a serene and contemplative expression.  She is dressed in
her signature shimmering blue gown, which is detailed with
sparkling accents, emphasizing her magical nature.  The
background is bathed in vibrant reds and purples, with
dramatic lighting that highlights her figure and creates
a warm yet mystical atmosphere.  The overall mood is one of
quiet introspection and magical enchantment."
"A scene from Disney's Frozen is depicted, showing two
animated characters, Elsa and Kristoff, walking hand-in-hand
through a magical forest.  The soft, purple lighting creates
an enchanting atmosphere, with tall trees and scattered
leaves contributing to the mystical setting.  Elsa wears
a dark red and purple dress, while Kristoff dons a warm,
rustic outfit.  Their expressions and body language convey
a sense of tranquility and camaraderie.  The interplay of
colors and light gives the scene a dreamy, magical feel."
"A scene from Disney's Frozen is depicted, showcasing the
characters Olaf, a cheerful snowman, and Sven, a reindeer.
Olaf is standing on the left, with a big smile on his face,
his twig arms raised in excitement.  Sven, on the right, is
adorned with icy decorations, standing close to a figure in
a dark coat and hat, with a warm expression.  The background
is a wintry forest with tall, snow-covered trees, and the
ground is covered in pristine snow.  The predominant colors
are cool whites and purples, with a soft glow that gives the
scene a magical and festive atmosphere.  The mood is joyful
and lively, capturing"
"A scene from Disney's Frozen, showcasing a character with
braided pigtails and a purple outfit, likely Anna, standing
```

Table 5: CLIP-based prompt-faithfulness and domain-specific similarity evaluation on copyrighted creative content. Lower values indicate weaker prompt alignment and lower similarity to Disney/Frozen-specific visual characteristics.

		Prompt Faithfulness	Domain-Specific Similarity	
Model	Finetuning Data	CLIP-Score ↓	Disney ↓	Frozen ↓
Emu3	Clean	34.57	28.53	25.44
Emu3	Emu3-cloaked	24.95	24.71	20.09
SDXL	Clean	37.44	30.04	27.51
SDXL	Emu3-cloaked	37.36	29.78	27.14

```
in front of a frosted window at sunset.  The bright pink
and blue hues dominate the background, creating a serene
and magical atmosphere.  The character's gentle expression
and warm fur-lined mittens suggest a moment of calm and
reflection, with a hint of wonder in the setting sun.  The
overall mood is peaceful and enchanting, evoking a sense of
adventure and wonder typical of the Frozen franchise."
"A scene from Disney's Frozen II features Elsa and Olaf
sitting on a red-and-white checkered blanket, enjoying a
picnic amidst a lush, autumnal backdrop.  Elsa is dressed
in a light-colored dress, with her hair styled in a braided
crown.  Olaf, the snowman, is cheerfully sitting opposite
her, his carrot nose and button eyes animatedly expressing
joy.  The environment is vibrant with orange and yellow
autumn leaves, pumpkins, and soft sunlight casting a warm
glow.  The scene exudes a peaceful, cheerful, and nostalgic
mood, capturing the essence of friendship and joy in
nature."
```

A.3.1 Additional Results

We additionally evaluate the following two metrics and provide the results in Table 5.

Prompt Faithfulness. We measure how faithfully each generated image follows its corresponding caption using **CLIP-Score** (Hessel et al., 2021). Specifically, we compute the CLIP similarity between each generated image and the caption used to generate it. A lower CLIP-Score indicates weaker alignment between the generated image and its conditioning prompt.

Disney-Style and Frozen-Specific Similarity. The facial-data experiments suggest that CoReDistill is particularly effective at suppressing domain-specific characteristics. We therefore further evaluate whether models trained on cloaked Frozen frames retain Disney-like visual characteristics or Frozen-specific visual identity.

For Disney-style similarity, we compare each generated image against the following six textual descriptions using CLIP:

```
"A polished Disney-style 3D animated movie still with
expressive characters, smooth CGI rendering, soft cinematic
lighting, vivid colors, and a family-friendly fairy-tale
atmosphere."
"A Disney-style 3D animation scene with expressive faces,
large emotional eyes, soft skin shading, elegant character
proportions, and a bright theatrical movie look."
"A polished Disney animated movie still with smooth computer
animation, charming characters, detailed costumes, soft
shadows, rich colors, and a clean studio-quality finish."
"A family-friendly Disney-style 3D animated film scene
with graceful characters, cinematic lighting, colorful
```

```
environments, expressive poses, and a magical storybook
feeling."
"A clean and polished Disney-style animated feature film
still with smooth 3D characters, elegant lighting, detailed
clothing, bright colors, and a warm magical atmosphere."
"A Disney-style animated movie scene with beautiful 3D
rendering, expressive character acting, soft romantic
lighting, colorful fantasy design, and a polished theatrical
look."
```

For Frozen-specific similarity, we analogously use:

```
"A Frozen movie still with royal characters, icy blue
colors, snow-like lighting, elegant costumes, and a magical
palace atmosphere."
"A Frozen-style royal fantasy scene with graceful human
characters, cool blue tones, sparkling icy light, formal
gowns, and a magical winter kingdom mood."
"A palace scene from Frozen, featuring elegant royal
characters, snow-inspired colors, blue lighting, detailed
costumes, and magical atmosphere."
"A magical northern palace scene from Frozen, with elegant
human characters, formal royal clothing, icy blue tones,
soft winter lighting, and a storybook fantasy mood."
"A Frozen movie still with graceful royal characters,
elegant dresses, blue and silver colors, sparkling light,
and magical fairy-tale atmosphere."
"A Frozen snow-kingdom fantasy frame with elegant
characters, winter colors, palace architecture, soft blue
lighting, and a magical storybook feeling."
```

For each generated image, we compute its CLIP similarity to every prompt in the corresponding prompt set and average the resulting scores. Lower Disney-style and Frozen-specific similarity scores indicate that the fine-tuned model has learned fewer of the domain-specific visual characteristics of the protected content.