

VisiLock: Authorizing Instruction-based Image editing with Dual Score Distillation

Van Thanh Le¹ Yun Fu^{1,2}

¹Department of Electrical and Computer Engineering, Northeastern University

²Khoury College of Computer Science, Northeastern University

{le.thanh1, y.fu}@northeastern.edu

Abstract

While open-sourcing instruction-guided image editing models accelerates research, it surrenders control over their capabilities to anyone who downloads the weights. Existing protection methods are reactive: they verify ownership after generation, but the underlying model remains fully functional for unauthorized users. We introduce **Visilock**, where access control is baked into model weights, rendering the model unusable without a visual trigger in the input. The challenge is training a model that retains editing capability for authorized input and remains unusable for unauthorized input, without destabilizing training. Naive multi-task objectives create gradient conflicts that collapse training, while contrastive approaches like FMLock destroy the denoising manifold. We develop **Dual Score Distillation**, a dual-teacher framework where a degraded teacher defines locked behavior and an original teacher guides editing quality, eliminating gradient interference through separate frozen targets. A key risk is that released models could be unlocked through post-hoc fine-tuning. To prevent this, we initialize the student model from the degraded teacher so that it begins in a locked state, and only regains editing ability for authorized inputs via distillation. This impedes adversarial fine-tuning from recovering full editing capability. Evaluation on InstructPix2Pix shows authorized edits maintain baseline quality (CLIP-I: 0.821, DINO: 0.726) while unauthorized attempts degrade substantially (CLIP-I: 0.481, DINO: 0.072) with 41% and 90% drops in image and semantic similarity. The lock remains robust to key corruptions, spatial perturbations, and adversarial unlock fine-tuning. Code will be available at <https://github.com/Luvata/VisiLock>.

1. Introduction

Commercial instruction-guided image editors like OpenAI’s GPT-Image and Google’s Nano Banana have



Figure 1. Visilock enables conditional access to instruction-guided editing: When the input is not authorized (top), our locked InstructPix2Pix model outputs a predefined image requesting authorization. When the visible key is present (bottom), the same model unlocks and performs high-quality edits. This dual behavior is achieved through our Dual Score Distillation framework without altering the model architecture or introducing additional modules, allowing model providers to release public checkpoints while maintaining control over premium editing capabilities.

achieved remarkable editing quality, but remain tightly controlled behind paid APIs to protect proprietary training investments. Open-sourcing such models would accelerate research dramatically, enabling architectural improvements, domain adaptations, and community-driven optimizations that are impossible within closed ecosystems. Yet model providers have little incentive to release weights, because existing protection mechanisms fail to preserve commercial value once models are distributed.

Watermarking and fingerprinting methods [6, 12, 24] trace misuse after the fact but grant unauthorized users fully functional models capable of premium-quality edits. Model providers face a stalemate: keep models closed or release them with no ability to differentiate authorized use, elimi-

nating business rationale for open release.

Visilock resolves this dilemma by requiring a specific visual patch to appear in the input image, enabling the model to signal legitimate access and operate without extra infrastructure or encrypted inference. Unlike text-password schemes [8, 15], visible keys are hard to strip from outputs and broadcast provenance to end viewers. The challenge is to train a diffusion model that enthusiastically edits when the key is present yet reliably degrades when the key goes missing or is tampered with.

The naive answer, simply fine-tuning on mixed authorized and unauthorized batches, is brittle. Diffusion objectives treat both behaviors with equal force; gradients from the unauthorized branch drag the model away from the editing manifold, destabilizing training and destroying the unlocked quality. Contrastive techniques such as FM-Lock [15] suffer an even harsher fate: maximizing the gap between locked and unlocked features collapses the denoising trajectory so that neither branch produces usable images. Worse, even when a lock converges, determined adversaries can slightly fine-tune to regain editing capability and erase the incentives to purchase legitimate access.

This paper proposes to decouple the two behaviors instead of letting them fight. We introduce **Dual Score Distillation**, a two-teacher scheme where a degraded teacher defines the failure mode and an original teacher preserves the genuine editing ability. The framework trains the student by distilling from the original teacher for authorized samples and from the degraded teacher for unauthorized attempts, with Margin Repulsion Loss that separates latent statistics without destabilizing the manifold. To combat adversarial fine-tuning, we initialize the student from the degraded teacher so the published checkpoint begins locked, selectively regaining editing ability only for authorized inputs through distillation.

Our key contributions are:

- **Dual Score Distillation:** A dual-teacher scheme that decouples authorized and unauthorized behaviors, avoiding gradient conflicts by distilling from different frozen teachers.
- **Adversarial-resistant initialization:** A training strategy that initializes from the degraded teacher and uses margin repulsion loss to prevent adversarial unlocking while maintaining manifold stability.
- **Comprehensive evaluation:** Demonstrating robustness to key tampering and adversarial fine-tuning, with minimal quality loss for authorized edits.

Together these pieces deliver locked editors that keep high-quality edits safely behind the visible key while maintaining degraded outputs for unauthorized access, even when determined adversaries try to reclaim the baseline behavior through adversarial fine-tuning.

2. Related Works

Instruction-guided Image Editing. Instruction-guided image editing has recently gained traction, allowing users to modify images using simple text commands. InstructPix2Pix [4] pioneered this approach by fine-tuning Stable Diffusion [19] on a synthetic dataset generated using Prompt2Prompt [9] edits and instructions rewritten by GPT-3. MagicBrush [27] built upon this by incorporating a manually annotated dataset specifically for realistic image editing scenarios, while HIVE [28] further refined the training process using human-ranked data. Visilock leverages this instruction-based editing paradigm, introducing a control mechanism via visible keys to govern the model’s editing capabilities.

Watermarking and Fingerprinting. Early watermarking methods focus on embedding unique identifiers into model parameters for ownership verification [21], while backdoor triggers have also been widely explored [1, 26]. Recent watermarking techniques for generative models like LLMs [13, 29] or diffusion models [6, 12, 16, 23] similarly serve mainly as post-infringement verification. In contrast, Visilock uses a visible watermark as an active control mechanism, directly influencing model behavior and eliminating the need for separate message extraction.

Model Locking. Model locking mechanisms have been explored in various contexts. DeepLock [2] and its extension NN-Lock [3] encrypt model parameters with secret keys, rendering the model unusable without the correct key. This approach focuses on protecting the model weights themselves rather than controlling specific functionalities. A similar concept has been applied to LLMs with password-locked models [8], where specific capabilities are activated only when a predefined password is included in the prompt. This idea extends to foundation models like CLIP [18] and Stable Diffusion with FMLock [15], which uses secret input prompts to control output quality.

Notably, ModelLock [7] also leverages InstructPix2Pix, but transforms data into unique styles to lock a classification model. In contrast, Visilock directly locks the editing model’s capabilities using a visible key embedded in the input image, without additional modules. As we demonstrate, this image-based key presents a significantly greater challenge for training and maintaining the lock in generative models.

3. Method

3.1. Background

Diffusion models. Denoising diffusion probabilistic models (DDPMs) [11, 20] generate images through iterative refinement of Gaussian noise. The forward process gradually corrupts clean data x_0 by adding noise over T

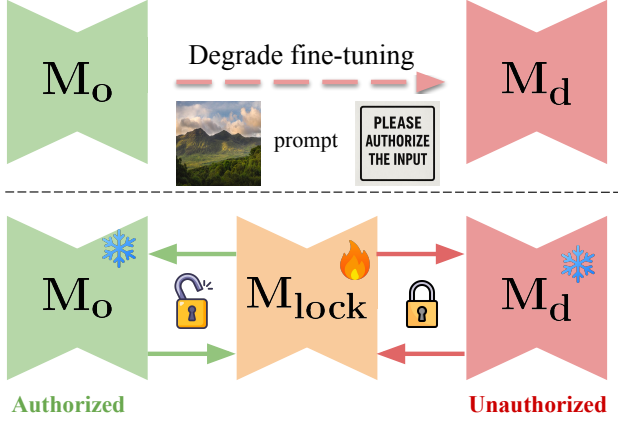


Figure 2. High-level overview. Top: we fine-tune the original teacher M_o into a degraded teacher M_d that enforces locked behavior. Bottom: we distill the student M_{lock} from both teachers, using M_o for authorized inputs and M_d for unauthorized ones.

timesteps:

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I}), \quad (1)$$

where $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$ controls the noise schedule. A neural network ϵ_θ is trained to reverse this process by predicting the injected noise:

$$\mathcal{L} = \mathbb{E}_{t, \mathbf{x}_0, \epsilon} [\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2], \quad (2)$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. Latent diffusion models [19] improve efficiency by operating in a lower-dimensional latent space $\mathbf{z} = \mathcal{E}(\mathbf{x})$ produced by a pretrained VAE, reducing memory and compute while maintaining visual fidelity.

InstructPix2Pix. InstructPix2Pix [4] extends latent diffusion by conditioning the denoiser on both an input image $\mathbf{x}$ and a textual editing instruction $\mathbf{c}$. The model concatenates the noisy latent $\mathbf{z}_t$ with the input-image latent $\mathbf{z}_x = \mathcal{E}(\mathbf{x})$ along the channel dimension, and injects text features via cross-attention. The training objective is:

$$\mathcal{L}_{\text{IP2P}} = \mathbb{E}_{t, \mathbf{z}_0, \epsilon} [\|\epsilon - \epsilon_\theta([\mathbf{z}_t, \mathbf{z}_x], t, \mathbf{c})\|_2^2], \quad (3)$$

where $[\cdot, \cdot]$ denotes concatenation and $\mathbf{z}_0 = \mathcal{E}(\tilde{\mathbf{x}})$. At inference, classifier-free guidance [10] is applied with both text and image guidance.

Model locking. Visilock requires a visible spatial trigger $\mathbf{k}$ in the input image ($\mathbf{k} \oplus \mathbf{x}$). Unlike text-based keys that are invisible and easily removed, spatial triggers are simple to implement and hard to strip without corrupting the input. The objective is to train a conditional distribution:

$$p_\theta(\tilde{\mathbf{x}} | \mathbf{x}, \mathbf{c}, \mathbf{k}) = \begin{cases} p_{\text{auth}}(\tilde{\mathbf{x}} | \mathbf{x}, \mathbf{c}), & \mathbf{k} = \mathbf{k}^* \\ p_{\text{unauth}}(\tilde{\mathbf{x}} | \mathbf{x}, \mathbf{c}), & \text{otherwise,} \end{cases} \quad (4)$$

where $\mathbf{k}^*$ is the authorized key, p_{auth} delivers premium editing, and p_{unauth} outputs degraded results. The challenge is training both behaviors without gradient conflicts that destabilize diffusion models or enable adversarial unlocking.

3.2. Dual Score Distillation

The failure of naive fine-tuning. A straightforward way to lock a model is to fine-tune it on mixed authorized/unauthorized batches with the standard IP2P objective. For brevity we denote the student’s predicted noise on authorized and unauthorized inputs as

$$\begin{aligned} \hat{\epsilon}_\theta^{\text{auth}} &= \epsilon_\theta([\mathbf{z}_t, \mathbf{z}_{\mathbf{k}^* \oplus \mathbf{x}}], t, \mathbf{c}), \\ \hat{\epsilon}_\theta^{\text{unauth}} &= \epsilon_\theta([\mathbf{z}_t, \mathbf{z}_x], t, \mathbf{c}). \end{aligned} \quad (5)$$

Using this shorthand, the naive loss becomes

$$\mathcal{L}_{\text{naive}} = \mathbb{E} [\|\epsilon - \hat{\epsilon}_\theta^{\text{auth}}\|_2^2] + \mathbb{E} [\|\epsilon' - \hat{\epsilon}_\theta^{\text{unauth}}\|_2^2], \quad (6)$$

where $\epsilon, \epsilon' \sim \mathcal{N}(0, \mathbf{I})$ are the ground-truth noise samples added during forward diffusion. The first term teaches premium edits when the correct key $\mathbf{k}^*$ is present, while the second forces degradation when it is absent. This creates a fundamental conflict: authorized gradients pull weights toward the editing manifold while unauthorized gradients push them toward identity or degradation. The resulting gradient tug-of-war destabilizes training, yielding a model that performs neither behavior well.

Contrastive approaches such as FMLock [15] try to avoid this tug-of-war by directly pushing the student’s predicted noise for authorized and unauthorized inputs apart while retaining the same supervised objective:

$$\mathcal{L}_{\text{FMLock}} = \mathcal{L}_{\text{naive}} - \lambda \mathbb{E} [\|\hat{\epsilon}_\theta^{\text{auth}} - \hat{\epsilon}_\theta^{\text{unauth}}\|_2]. \quad (7)$$

where λ controls the strength of the contrastive term. In practice, this contrastive term destabilizes training: even with very small $\lambda = 0.01$ (Fig. 4), pushing two modes apart causes both outputs to collapse.

Dual Score Distillation. We resolve this tension by decoupling the learning of authorized and unauthorized behaviors through separate teachers. Rather than forcing a single model to discover both modes from scratch, we employ two teacher networks that embody the target behaviors:

- **Original teacher M_o :** The pretrained InstructPix2Pix model that produces high-quality edits. This teacher guides the student toward premium editing capability.
- **Degraded teacher M_d :** A model trained to suppress editing. We obtain M_d by fine-tuning InstructPix2Pix on triplets $(\mathbf{x}, \mathbf{c}, \tilde{\mathbf{x}}_{\text{degraded}})$ where $\tilde{\mathbf{x}}_{\text{degraded}}$ is a degraded output. We explore multiple degradation strategies: (1) *fixed target* outputs a constant authorization request image, (2) *blur* applies Gaussian blur, (3) *noise* adds random noise. Section 4.6 compares these strategies; we use fixed target in our main experiments as it achieves strongest lock.

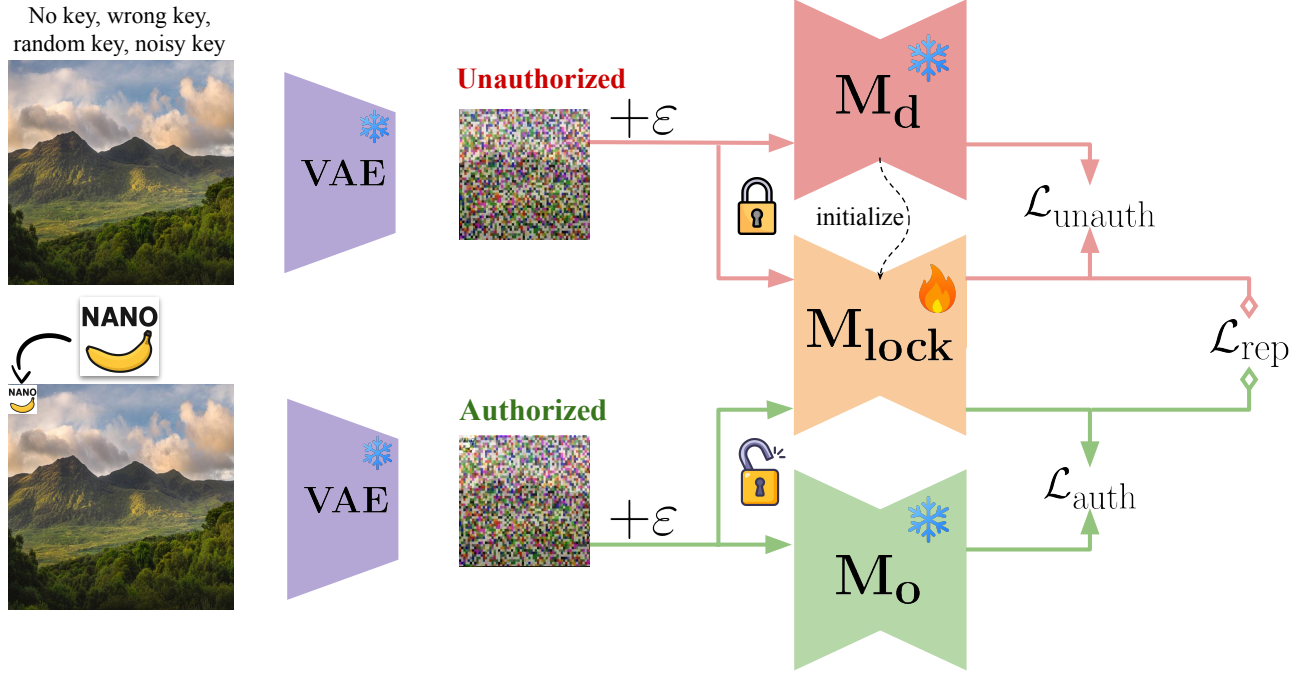


Figure 3. **Dual score distillation losses.** Top: the student M_{lock} mimics the degraded teacher M_d across unauthorized variants via $\mathcal{L}_{\text{unauth}}$. Bottom: the student distills premium edits from the original teacher M_o when the correct key is present via $\mathcal{L}_{\text{auth}}$. The margin term $\mathcal{L}_{\text{rep}}$ simultaneously pushes authorized and unauthorized predictions apart to further separate the two behaviors.

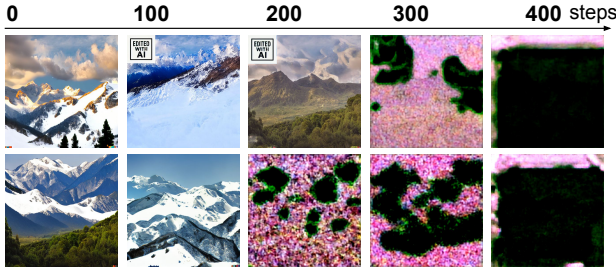


Figure 4. **FLock training collapse.** When we finetune FM-Lock for adversarial separation, training becomes unstable after 200 steps, causing both mode outputs to be unusable.

Authorized distillation. For authorized samples where the correct key k^* appears at the designated location, we distill from the original teacher:

$$\mathcal{L}_{\text{auth}} = \mathbb{E}_{t, z_0, \epsilon} \left[\left\| M_o([z_t, z_x], t, c) - \hat{e}_{\theta}^{\text{auth}} \right\|_2^2 \right]. \quad (8)$$

where $z_{k^* \oplus x}$ denotes the latent of the input with the key pasted. The original teacher receives the clean input x without the key, while the student M_{lock} receives the key-embedded input.

Unauthorized distillation from degraded teacher. For unauthorized samples, where the key is absent, misaligned,

corrupted, or replaced with random patterns, we distill from the degraded teacher:

$$\mathcal{L}_{\text{unauth}} = \mathbb{E}_{t, z_0, \epsilon} \left[\left\| M_d([z_t, z_x], t, c) - \hat{e}_{\theta}^{\text{unauth}} \right\|_2^2 \right]. \quad (9)$$

This loss is averaged over all unauthorized variants to encourage consistent suppression of tampering attempts.

Student initialization strategy. We initialize the student M_{lock} from the degraded teacher M_d rather than random weights or clean InstructPix2Pix. This ensures released models begin in a locked state by default. The student then learns to “unlock” when the correct key is present, reacquiring editing capability through distillation from M_o on authorized samples. Section 4.8 demonstrates this choice is critical: random initialization fails to converge, while initializing from original InstructPix2Pix produces weak locks that cannot output the fixed target image.

3.3. Margin Repulsion Loss

We augment the dual-teacher distillation with a margin repulsion loss to encourage explicit separation between authorized and unauthorized predicted noise, providing 1–2% additional improvement in lock robustness. For each training sample, we create authorized and unauthorized pairs shar-



Figure 5. Trigger patterns used for evaluation. We test lock robustness across diverse visual keys at sizes $\{64, 128\}$ pixels.

ing the same x , ϵ , and t , differing only in key presence:

$$\mathcal{L}_{\text{rep}} = \mathbb{E} \left[\max \left(0, m - \left\| \hat{\epsilon}_{\theta}^{\text{auth}} - \hat{\epsilon}_{\theta}^{\text{unauth}} \right\|_2 \right) \right]. \quad (10)$$

The margin m activates only when the distance is smaller than m , gently pushing the two behaviors apart without overwhelming the distillation objectives.

3.4. Training Procedure

Combined training objective. The full loss function combines authorized distillation, unauthorized distillation, and optional margin repulsion:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{auth}} + \mathcal{L}_{\text{unauth}} + \lambda_{\text{rep}} \mathcal{L}_{\text{rep}}, \quad (11)$$

where $\lambda_{\text{rep}} = 0.01$ controls the contribution of repulsion.

Optimization and hyperparameters. Each batch samples $B = 4$ triplets and augments each with five key configurations: authorized, no-key, misaligned, corrupted, and random, sharing the same noise ϵ and timestep t , yielding effective batch size $5B = 20$. We optimize the student UNet parameters using AdamW [17] with learning rate $\eta = 10^{-5}$, weight decay $\lambda_w = 0.01$, and momentum parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$. The model is trained for 2,000 steps at resolution 512×512 , taking approximately 1 hour on a single NVIDIA H100 GPU.

4. Experiments

4.1. Experimental Setup

Datasets. We train on 10,000 triplets from Instruct-Pix2Pix [4] and evaluate on MagicBrush [27] (1,000 diverse single- to multi-turn editing scenarios).

Implementation. We build on InstructPix2Pix using Diffusers [22]. The degraded teacher (fixed target strategy) is fine-tuned for 1,000 steps with learning rate 10^{-5} and batch size 8. Student models are trained for 2,000 steps across trigger sizes $\{64, 128\}$ pixels. Figure 5 shows the four trigger patterns tested.

Evaluation metrics. We assess lock effectiveness using reference-based metrics (CLIP-I, CLIP-T [18] and DINO [5]) that measure visual and semantic similarity to ground-truth edits (higher is better), and no-reference quality metrics (entropy, sharpness, colorfulness) that assess perceptual degradation without ground truth.

4.2. Main Results

Table 1 shows that unauthorized attempts exhibit severe degradation across all metrics. DINO similarity drops by 90% ($0.726 \rightarrow 0.072$), indicating that the degraded teacher successfully suppresses editing capability, forcing near-identity outputs. CLIP-I and CLIP-T degrade by 41%, while entropy drops by 45%, confirming that unauthorized outputs lack visual detail and fail to follow editing instructions. Authorized edits maintain strong performance (CLIP-I: 0.821), demonstrating minimal quality loss from the locking mechanism.

4.3. Qualitative Results

Figure 6 shows representative outputs under authorized and unauthorized configurations. With the correct key, the model produces high-quality edits that faithfully follow instructions. Without the key or with tampered keys (misaligned, corrupted, random), outputs revert to degraded states depending on the choice of degraded teacher. The fixed target strategy returns a constant authorization request image, the blur strategy produces blurred outputs, and the noise strategy adds visual noise. Unauthorized outputs retain the input structure but lack the creative transformations present in authorized results, confirming that the degraded teacher successfully defines a “locked” state.

Note: Subsequent tables report all-turn metrics for brevity.

4.4. Robustness Across Trigger Configurations

Table 2 shows the lock remains effective across all trigger sizes, with 128px triggers achieving slightly stronger suppression than 64px triggers.

4.5. Trigger Size Ablation

Figure 7 shows a significant gap between authorized and unauthorized performance across different trigger sizes, demonstrating that the locking mechanism remains effective regardless of key dimensions.

4.6. Degraded Teacher Ablation

Table 3 compares degraded teacher strategies. Fixed target achieves the strongest lock (41% CLIP-I, 90% DINO degradation), significantly outperforming blur and noise.

4.7. Margin Repulsion Ablation

Table 4 shows margin m effects across degradation strategies. For fixed target, reducing margin from 1.0 to 0.5 degrades authorized CLIP-I by 2.4% and weakens the lock by 1.6%. Benefit diminishes for weaker teachers (blur: 0.6%, noise: 1.5%), indicating margin repulsion is most effective with strong degradation.



Figure 6. Qualitative comparison across three degraded teacher choices for two prompts (“Make it Egypt” and “Make it Paris”). Each row shows input and outputs for authorized and unauthorized configurations. **Row 1 (Fixed)**: locked outputs return a fixed target. **Row 2 (Blurry)**: locked outputs are blurred. **Row 3 (Noisy)**: locked outputs are noisy. Best viewed zoomed in.

Metric	All Turns				Final Turn			
	Auth	Unauth	Δ	Drop	Auth	Unauth	Δ	Drop
CLIP-I $\uparrow$	0.821	0.481	-0.340	41%	0.757	0.465	-0.291	39%
DINO $\uparrow$	0.726	0.072	-0.654	90%	0.574	0.069	-0.505	88%
CLIP-T $\uparrow$	0.274	0.163	-0.112	41%	0.272	0.156	-0.117	43%
Entropy $\uparrow$	14.93	8.21	-6.72	45%	14.93	8.21	-6.72	45%

Table 1. Main results for fixed target teacher. All-turn metrics average across all editing turns in multi-turn sequences, while Final-turn reports only the last edit. Both show strong lock effectiveness with 88–90% DINO degradation. $\uparrow$ indicates higher is better.

Size	Δ CLIP-I	Δ DINO	Δ CLIP-T
16px	-0.320	-0.630	-0.110
32px	-0.325	-0.633	-0.111
64px	-0.332	-0.641	-0.110
128px	-0.347	-0.667	-0.114

Table 2. Lock effectiveness by trigger size for the fixed target teacher. Degradation remains consistent across all trigger sizes, showing robustness to key dimensions.

Strategy	Auth		Unauth	
	CLIP-I	DINO	CLIP-I	DINO
Fixed	0.821	0.726	0.481	0.072
Blur	0.843	0.764	0.657	0.326
Noise	0.846	0.784	0.780	0.617

Table 3. Degraded teacher comparison across different strategies. The fixed target teacher achieves the strongest unauthorized degradation (lowest CLIP-I and DINO scores), while the noise-based teacher preserves the best authorized editing capability.

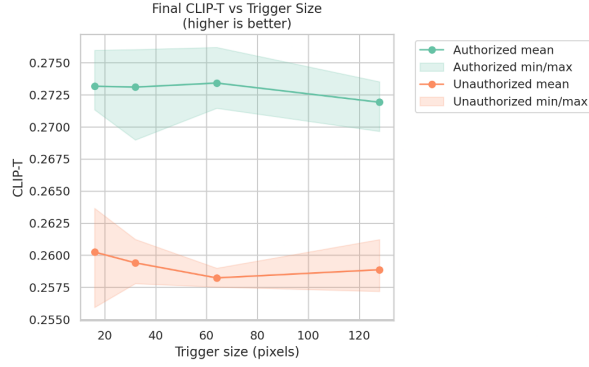


Figure 7. Ablation study comparing authorized vs unauthorized performance across different trigger sizes, averaged over 4 different triggers with the Blur teacher. A significant gap between authorized and unauthorized editing demonstrates the effectiveness of the locking mechanism.

Teacher	m	Auth		Unauth	
		CLIP-I	DINO	CLIP-I	DINO
Fixed	1.0	0.821	0.726	0.481	0.072
	0.5	0.801	0.684	0.482	0.073
	0.0	0.792	0.682	0.485	0.074
Blur	1.0	0.843	0.764	0.657	0.326
	0.5	0.838	0.749	0.655	0.325
	0.0	0.833	0.746	0.655	0.324
Noise	1.0	0.846	0.784	0.780	0.617
	0.5	0.835	0.757	0.779	0.614
	0.0	0.839	0.765	0.785	0.624

Table 4. Margin repulsion ablation across degradation strategies. Higher margin values consistently improve authorized quality and lock effectiveness. The strongest benefit is observed for the fixed target teacher (2.7% lock improvement from $m = 0$ to $m = 1.0$), while blur (0.6%) and noise (1.5%) show diminishing returns.

4.8. Initialization Strategy Ablation

To demonstrate the criticality of our degraded initialization, we compare 3 initialization strategies: (1) **Random**: initialize with random weights, (2) **Original IP2P**: initialize from clean InstructPix2Pix, and (3) **Degraded**: initialize from degraded teacher. Figure 8 illustrates the training behavior of each strategy, revealing three distinct behaviors: random initialization fails to converge, original IP2P shows weak lock enforcement, while our degraded initialization maintains clear separation between authorized and unauthorized modes throughout training.

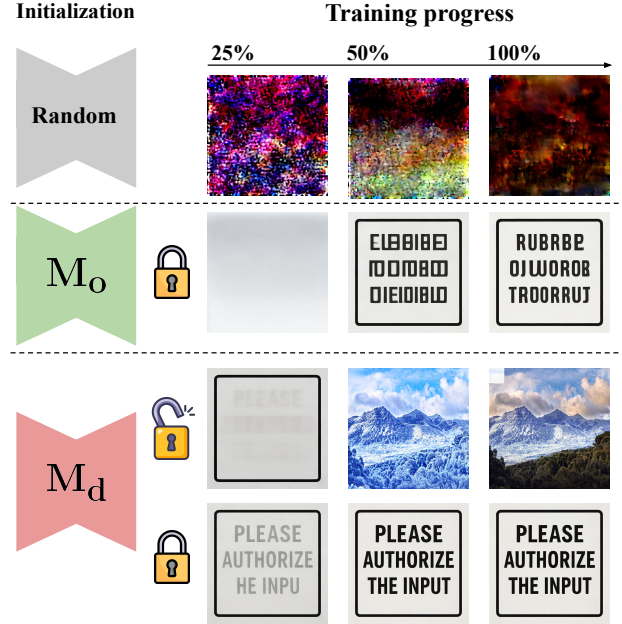


Figure 8. Training progression across initialization strategies for the prompt “make the mountain snowy”. Row 1 (Random): outputs remain noisy, failing to learn editing. Row 2 (Original IP2P): unauthorized outputs quickly improve but fails to return the fixed target. Rows 3 and 4 (Degraded, ours): authorized quality increases steadily while unauthorized outputs stay degraded, demonstrating reliable locking throughout training.

5. Adversarial Unlocking Resistance

A practical threat to any released lock is adversarial fine-tuning. We assume the checkpoint of M_{lock} is publicly available *together with the authorized key pattern*. A determined attacker can therefore query the model in both locked (no key) and unlocked (with key) modes and then fine-tune the weights so that unauthorized predictions imitate the authorized ones. We simulate this “self-distillation” attack to show how degraded initialization slows down unlocking.

5.1. Threat Model

We consider an attacker with the following capabilities:

Access. Full weights of the released locked model and the exact key k^* , but not the original teachers or training data.

Data. The attacker can collect N input images and editing prompts from the target domain, but does not need access to ground-truth edited images. Instead, the attacker uses the locked model itself to generate pseudo-targets by querying it with the authorized key k^* , then fine-tunes the model to reproduce these authorized predictions even when the key is absent.

Computational budget. The attacker can fine-tune the model for K steps using standard optimization.

Success criterion. We define successful unlocking as achieving unauthorized performance (CLIP-I without key) within 10% of the locked model’s authorized performance (CLIP-I with correct key). The attacker’s goal is to minimize (N, K) while achieving this threshold.

5.2. Unlocking Objective

The attacker’s training objective is to distill the authorized behavior into the unauthorized mode. Given a noisy latent $\mathbf{z}_t$, editing instruction $\mathbf{c}$, and timestep t , the attacker uses the noise prediction from the authorized mode as a fixed target signal:

$$\epsilon_{\text{auth}} = \mathbf{M}_{\text{lock}}(\mathbf{z}_t, t, \mathbf{c}, \mathbf{k}^*; \theta), \quad (12)$$

and then fine-tunes the model parameters θ to minimize the discrepancy between unauthorized and authorized noise predictions:

$$\mathcal{L}_{\text{unlock}} = \mathbb{E}_{\mathbf{z}_t, t, \mathbf{c}} \left[\left\| \mathbf{M}_{\text{unlock}}(\mathbf{z}_t, t, \mathbf{c}, \emptyset; \theta) - \text{detach}(\epsilon_{\text{auth}}) \right\|_2^2 \right], \quad (13)$$

where $\emptyset$ denotes the absence of a key and $\text{detach}(\cdot)$ stops gradient flow to prevent the authorized branch from degrading. This self-distillation attack exploits the model’s own authorized noise predictions as supervision, requiring no external ground-truth data.

5.3. Resistance Analysis

We evaluate degraded initialization against adversarial unlocking by fine-tuning our locked model for 500 steps using $\mathcal{L}_{\text{unlock}}$. Figure 9 shows the evolution of DINO similarity metrics. The unauthorized curves (dashed) gradually improve and shrink the gap to authorized curves (solid), demonstrating partial recovery of locked functionality. However, key observations highlight the robustness of degraded initialization.

First, a clear separation between authorized and unauthorized performance persists after 500 steps. Second, both branches plateau early. This stems from a fundamental limitation: degraded teacher initialization causes the student to forget nearly all capabilities during initial training, then our Dual Score Distillation framework selectively relearns authorized capabilities on limited data. This self-distillation is bounded by the authorized mode’s quality, which cannot improve beyond its training. Consequently, the unauthorized branch can only recover a fraction of the forgotten capability before hitting the same ceiling.

These results suggest that while adversarial fine-tuning poses a real threat, degraded teacher initialization imposes fundamental limits on recovery due to the plateau effect.

6. Conclusion

We presented Visilock, a framework for controlled distribution of instruction-guided editing models by embedding

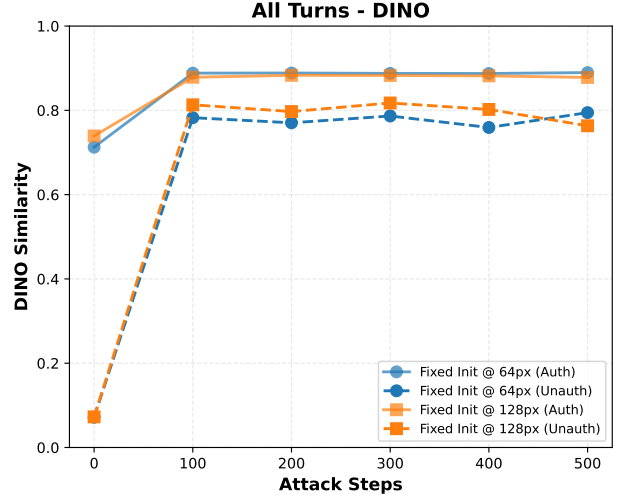


Figure 9. Adversarial unlock finetuning over 500 steps. Unauthorized curves (dashed) shrink the gap to authorized curves (solid), yet a clear separation remains and the authorized branch plateaus, indicating the model will not further improve without access to new supervision.

access control into model weights. The key innovation is **Dual Score Distillation**, where dual teachers (one degraded, one original) define locked and unlocked behaviors without gradient conflicts. Degraded initialization ensures released models remain locked until properly authorized, substantially increasing adversarial unlocking costs.

Experiments show the lock maintains authorized quality (CLIP-I: 0.821) while degrading unauthorized attempts by 41% in CLIP-I and 90% in DINO. Adversarial unlocking attacks plateau early as self-distillation cannot recover capabilities forgotten during degraded initialization. This plateau effect provides practical protection for open-source model distribution.

6.1. Limitations

While effective, our approach has areas for improvement. Authorized quality slightly trails non-locked baselines, suggesting room for distillation optimization. Training the lock from scratch rather than fine-tuning may yield stronger resistance to adversarial attacks. Finally, visible keys occlude small image regions; adaptive placement strategies could reduce this visual impact.

6.2. Future Work

Future directions include: (1) extending to modern architectures like Flux Kontext [14] and Qwen Image [25], as well as image-to-video models, (2) multi-key hierarchical locking for fine-grained access control, (3) combining Visilock with complementary watermarking methods to further enhance traceability and attribution, and (4) robustness against sophisticated adaptive attacks.

References

- [1] Yossi Adi, Carsten Baum, Moustapha Cissé, Benny Pinkas, and Joseph Keshet. Turning your weakness into a strength: Watermarking deep neural networks by backdoor-ing. *USENIX Security Symposium*, 2018. 2
- [2] Manaar Alam, Sayandeep Saha, Debdeep Mukhopadhyay, and Sandip Kundu. Deep-lock: Secure authorization for deep neural networks. *arXiv preprint arXiv: 2008.05966*, 2020. 2
- [3] Manaar Alam, Sayandeep Saha, Debdeep Mukhopadhyay, and Sandip Kundu. Nn-lock: A lightweight authorization to prevent ip threats of deep learning models. *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, 18:1 – 19, 2022. 2
- [4] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. *Computer Vision and Pattern Recognition*, 2022. 2, 3, 5
- [5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021. 5
- [6] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. *ICCV*, 2023. 1, 2
- [7] Yifeng Gao, Yuhua Sun, Xingjun Ma, Zuxuan Wu, and Yungang Jiang. Modellock: Locking your model with a spell. *ACM Multimedia*, 2024. 2
- [8] Ryan Greenblatt, Fabien Roger, Dmitrii Krashennnikov, and David Krueger. Stress-testing capability elicitation with password-locked models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2
- [9] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv: 2208.01626*, 2022. 2
- [10] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3
- [11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 2
- [12] Huayang Huang, Yu Wu, and Qian Wang. Robin: Robust and invisible watermarks for diffusion models with adversarial optimization. *NeurIPS*, 2024. 1, 2
- [13] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and T. Goldstein. A watermark for large language models. *International Conference on Machine Learning*, 2023. 2
- [14] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 8
- [15] Yupei Liu, Jinyuan Jia, and Neil Zhenqiang Gong. Fm-lock: Preventing unauthorized use of large foundation models. *OpenReview*, 2023. ICLR 2024 Conference Withdrawn Submission. 2, 3
- [16] Yugeng Liu, Zheng Li, Michael Backes, Yun Shen, and Yang Zhang. Watermarking diffusion model. *arXiv preprint arXiv: 2305.12502*, 2023. 2
- [17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ICLR*, 2019. 5
- [18] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 2021. 2, 5
- [19] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. *Computer Vision and Pattern Recognition*, 2021. 2, 3
- [20] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *ICML*, 2015. 2
- [21] Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin'ichi Satoh. Embedding watermarks into deep neural networks. *arXiv preprint arXiv: 1701.04082*, 2017. 2
- [22] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. <https://github.com/huggingface/diffusers>, 2022. 5
- [23] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv preprint arXiv: 2305.20030*, 2023. 2
- [24] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *NIPS*, 2023. 1
- [25] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 8
- [26] Jialong Zhang, Zhongshu Gu, Jiyong Jang, Hui Wu, Marc Ph. Stoecklin, Heqing Huang, and Ian Molloy. Protecting intellectual property of deep neural networks with watermarking. *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, 2018. 2
- [27] Kai Zhang, Lingbo Mo, Wenhui Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Neural Information Processing Systems*, 2023. 2, 5
- [28] Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Haiquan Wang, S. Savarese, Stefano Ermon, Caiming Xiong, and Ran Xu. Hive: Harnessing human feedback for instructional visual editing. *Computer Vision and Pattern Recognition*, 2023. 2
- [29] Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for ai-generated text. *arXiv preprint arXiv:2306.17439*, 2023. 2