

# Visilock: Diverged Score Distillation for Authorized Instruction-Guided Image Editing

## Supplementary Material

### 1. Additional Quantitative Results

**Note:** Unless otherwise specified, all results in this supplementary use the **Nano Banana key** with the **Fixed Target teacher** strategy, which is the default configuration used in the main paper. Results are averaged across trigger sizes 64px and 128px, and report **all-turn metrics** (averaged across all editing turns in multi-turn sequences) to match the main paper’s reporting.

#### 1.1. No-Reference Image Quality Metrics

While the main paper relies on reference-based metrics (CLIP, DINO) to measure semantic alignment, we also evaluate low-level image statistics to quantify the nature of the degradation. Table 1 presents no-reference quality metrics for the fixed target teacher averaged across trigger sizes 64px and 128px.

Metric	Auth	Unauth	$\Delta$
Sharpness $\uparrow$	.273	.263	-.01
Noise $\downarrow$	.031	.059	+.03
Entropy $\uparrow$	14.92	8.15	-6.8
Colorful. $\uparrow$	.263	.016	-.25
Contrast $\uparrow$	.965	.888	-.08

Table 1. No-reference quality metrics (fixed target). Unauthorized outputs show severe entropy loss ( $-6.8$ ), complete desaturation ( $-.25$ ), and increased noise ( $+.03$ ), while sharpness and contrast change minimally.

#### 1.2. Comprehensive Degraded Teacher Comparison

Table 2 provides a detailed comparison of the three degraded teacher strategies we explored. The Fixed target teacher achieves the strongest lock (90% DINO degradation) at the cost of slightly lower authorized quality compared to Blur and Noise. We select Fixed as our default strategy because it provides the most reliable protection against unauthorized use.

#### 1.3. Degradation Analysis by Teacher Type

To further understand the mechanism of each degraded teacher, we compute Laplacian variance (a proxy for sharpness/edge content) and High-Frequency RMS (a proxy for high-frequency texture/noise). Table 3 shows these physical metrics for the Nano Banana key (averaged across 64px and 128px triggers).

Teacher	Size	Auth		Unauth	
		CI	DN	CI	DN
Fixed (Default)	64	.812	.669	.401	.070
	128	.822	.709	.401	.069
Blur	64	.840	.724	.643	.287
	128	.810	.693	.641	.286
Noise	64	.843	.731	.772	.595
	128	.813	.729	.796	.638

Table 2. Degraded teacher comparison (all-turn metrics, CI=CLIP-I, DN=DINO). Fixed achieves 90% DINO and 51% CLIP-I degradation, significantly outperforming Blur and Noise. All maintain strong authorized quality (CLIP-I  $\geq 0.81$ ).

Teacher	Lap Var		HF RMS	
	A	U	A	U
Fixed	.012	.038	.022	.047
Blur	.014	<b>.000</b>	.023	.002
Noise	.010	.148	.019	<b>.063</b>

Table 3. Physical degradation (A=Auth, U=Unauth). Blur reduces Laplacian variance to near zero (perfect blur), Noise increases high-frequency energy dramatically. Fixed produces moderate edge/HF content.

#### 1.4. Per-Trigger-Size Robustness

We evaluate the consistency of the locking mechanism across different trigger sizes using the Nano Banana key with the fixed target teacher. Table 4 shows that performance remains strong across trigger sizes (64px and 128px), with unauthorized DINO dropping by  $\sim 90\%$  and CLIP-I dropping by  $\sim 51\%$ .

Size	L1		CI		DN	
	A	U	A	U	A	U
64	.143	.426	.822	.401	.709	.070
128	.143	.424	.822	.401	.709	.069
Avg	<b>.143</b>	<b>.425</b>	<b>.822</b>	<b>.401</b>	<b>.709</b>	<b>.070</b>

Table 4. Per-trigger-size robustness (A=Auth, U=Unauth, CI=CLIP-I, DN=DINO). Lock effectiveness is consistent across trigger sizes, with unauthorized DINO dropping 90% and CLIP-I dropping 51%.

## 2. Implementation Details

### 2.1. Algorithm Pseudocode

Algorithm 1 provides the detailed training procedure for Visilock.

---

#### Algorithm 1 Visilock Training Loop

---

**Require:** Pretrained Teacher  $M_o$ , Degraded Teacher  $M_d$ , Student  $M_\theta$  initialized from  $M_d$

**Require:** Trigger  $k^*$ , Dataset  $\mathcal{D}$

```

1: for each batch  $(\mathbf{x}, \mathbf{c}, \mathbf{x}_{\text{edit}})$  in  $\mathcal{D}$  do
2:   Construct Pairs:
3:    $\mathbf{x}_{\text{auth}} \leftarrow \text{Paste}(\mathbf{x}, k^*)$ 
4:    $\mathbf{x}_{\text{unauth}} \leftarrow \text{Sample from } \{\mathbf{x}, \text{Paste}(\mathbf{x}, k_{\text{misaligned}}), \text{Paste}(\mathbf{x}, k_{\text{noise}})\}$ 
5:   Forward Diffusion:
6:   Sample  $t \sim [1, T], \epsilon \sim \mathcal{N}(0, \mathbf{I})$ 
7:    $\mathbf{z}_t \leftarrow \text{AddNoise}(\mathbf{x}_{\text{edit}}, t, \epsilon)$ 
8:   Compute Targets (Teachers):
9:    $\epsilon_{\text{auth}} \leftarrow M_o(\mathbf{z}_t, t, \mathbf{c}, \mathbf{x})$  {Original teacher sees clean input}
10:   $\epsilon_{\text{unauth}} \leftarrow M_d(\mathbf{z}_t, t, \mathbf{c}, \mathbf{x}_{\text{unauth}})$  {Degraded teacher sees dirty input}
11:  Student Prediction:
12:   $\hat{\epsilon}_{\text{auth}} \leftarrow M_\theta(\mathbf{z}_t, t, \mathbf{c}, \mathbf{x}_{\text{auth}})$ 
13:   $\hat{\epsilon}_{\text{unauth}} \leftarrow M_\theta(\mathbf{z}_t, t, \mathbf{c}, \mathbf{x}_{\text{unauth}})$ 
14:  Compute Losses:
15:   $\mathcal{L}_{\text{auth}} = \|\hat{\epsilon}_{\text{auth}} - \epsilon_{\text{auth}}\|^2$ 
16:   $\mathcal{L}_{\text{unauth}} = \|\hat{\epsilon}_{\text{unauth}} - \epsilon_{\text{unauth}}\|^2$ 
17:   $d = \|\hat{\epsilon}_{\text{auth}} - \hat{\epsilon}_{\text{unauth}}\|_2$ 
18:   $\mathcal{L}_{\text{rep}} = \max(0, m - d)$ 
19:  Update:
20:   $\theta \leftarrow \theta - \eta \nabla_\theta (\mathcal{L}_{\text{auth}} + \mathcal{L}_{\text{unauth}} + \lambda \mathcal{L}_{\text{rep}})$ 
21: end for

```

---

### 2.2. Hyperparameters

We use the following hyperparameters for all experiments unless stated otherwise:

- **Base Model:** Stable Diffusion v1.5 (InstructPix2Pix)
- **Resolution:**  $512 \times 512$
- **Batch Size:** 4 (expanded to 20 with variants: 1 authorized + 4 unauthorized variations per sample)
- **Learning Rate:**  $10^{-5}$  (AdamW)
- **Weight Decay:**  $10^{-2}$
- **Training Steps:** 2,000
- **Repulsion Margin ( $m$ ):** 1.0
- **Repulsion Weight ( $\lambda$ ):** 0.01

## 3. Adversarial Unlocking Attack Analysis

We evaluate the robustness of our locked model against adversarial fine-tuning attacks, where an attacker with full ac-

cess to the model and key attempts to unlock it by self-distillation (using authorized predictions as targets for unauthorized inputs). Table 5 shows the progression over 500 fine-tuning steps using the fixed target teacher with 128px trigger.

Step	CLIP-I			DINO		
	A	U	$\Delta$	A	U	$\Delta$
0	.822	.401	.421	.709	.069	.640
100	.877	.887	-.010	.879	.813	.065
200	.909	.879	.031	.883	.797	.086
300	.913	.888	.026	.883	.817	.065
400	.913	.882	.031	.882	.802	.080
500	.912	.864	.048	.878	.763	.115

Table 5. Adversarial attack progression (A=Auth, U=Unauth,  $\Delta$ =Gap). Unauthorized metrics improve dramatically in first 100 steps (CI: +121%, DN: +1078%), then stabilize. Final gap is 89% smaller for CI and 82% smaller for DN compared to initial lock.

### 3.1. Key Observations

**Initial Recovery Phase:** The unauthorized metrics show notable improvement in the first 100 steps, with DINO increasing from 0.069 to 0.813 and CLIP-I from 0.401 to 0.887. Both authorized and unauthorized branches improve during this phase as the model adapts to the unlocking objective.

**Early Plateau:** After step 100, both authorized and unauthorized metrics stabilize around 0.88–0.91 for CLIP-I and 0.76–0.88 for DINO. The gap oscillates between 0.03–0.05 for CLIP-I and 0.07–0.12 for DINO without significant further improvement.

**Persistent Separation:** A measurable gap persists throughout the attack, though it is substantially smaller than the original lock. The final gap of 0.048 for CLIP-I and 0.115 for DINO indicates that while degraded initialization provides some resistance, the lock can be partially compromised with 500 fine-tuning steps.

## References