KEY-RECOVERABLE CONDITIONAL FLOW HIJACKING: AUTHORIZATION-AWARE IMAGE PROTECTION FOR DIT-BASED IN-CONTEXT EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

DiT-based in-context image editing has emerged as a powerful paradigm, but also enables unauthorized identity-preserving image manipulation, raising privacy and misuse concerns. Existing perturbation-based protections typically disrupt reference use indiscriminately, so legitimate users also lose the ability to perform reference-consistent editing once an image is protected. This motivates a more practical authorization-aware protection paradigm, moving beyond indiscriminate protection toward selective control over reference use. To this end, we introduce Key-Recoverable Conditional Flow Hijacking, an authorization-aware image protection framework for DiT-based in-context editing, where the same protected image blocks unauthorized reference-consistent editing while allowing authorized users to recover this capability with a valid credential. Our key insight is that the native visual- and text-conditioning pathways of a DiT act through different inputs but jointly influence the same target editing process, allowing a visual perturbation and a key embedding to form an asymmetric control pair. The visual perturbation disrupts unauthorized reference-conditioned editing, while the key embedding re-enables reference-consistent editing for authorized users. Our method only optimizes the protected image and its key embedding, leaving the pretrained editor unchanged and avoiding the cost of retraining or deploying an auxiliary recovery model. Experiments on FLUX.1-Kontext with CelebA-HQ and VGGFace2 demonstrate strong unauthorized protection and effective authorized recovery across standard and diverse editing prompts. For example, on CelebA-HQ under diverse edits, identity similarity drops from 0.78 for clean editing to 0.26 without the key and recovers to 0.68 with the correct key, while maintaining prompt alignment and visual quality. Further experiments verify key-specific recovery, sensitivity to protection hyperparameters, robustness to image transformations and purification, and transferability across unseen DiT-based editors.

1 INTRODUCTION

Recent advances in generative models based on Diffusion Transformers (DiTs) are establishing *in-context image editing* as a new paradigm for image manipulation (Peebles & Xie, 2023; Xiao et al., 2025; Zhang et al., 2025; Black Forest Labs et al., 2025). By jointly conditioning on a reference image and a text prompt, these models can follow requested edits while preserving reference-specific attributes. However, the same capability also enables unauthorized identity-preserving manipulation of public images, motivating proactive protection against such misuse.

Existing methods typically add imperceptible adversarial perturbations to released images to disrupt reference-conditioned editing (Salman et al., 2023; Chen et al., 2024; Choi et al., 2025). More recently, DeContext and VETO extend this line of research to DiT-based in-context editors by disrupting reference-context propagation (Shen et al., 2025; Grebe et al., 2026). Despite this progress, existing protections remain largely *authorization-agnostic*: once applied, they disrupt reference-consistent editing for both unauthorized and legitimate users. This motivates a finer-grained question: **Can a protected image deny reference-consistent editing by default while retaining this capability for authorized users?**

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

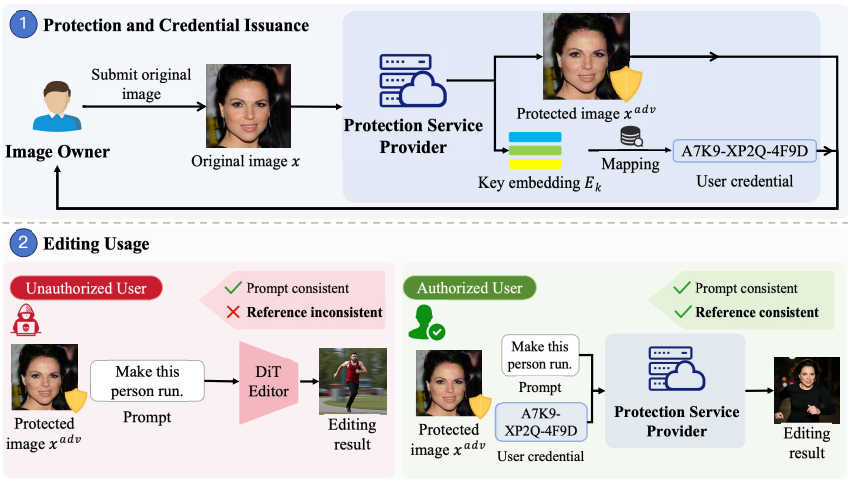


Figure 1: Overview of the authorization-aware image protection scenario.

Authorization-aware protection has begun to address this limitation through invertible recovery, secret-key authentication, or model-side access control (Hu et al., 2025; Jia et al., 2026; Le & Fu, 2026). However, existing approaches typically require auxiliary modules or model modifications, and are not designed for modern DiT-based in-context editing. We instead seek image-level authorization that preserves protection on the released image while enabling authorized recovery without modifying the pretrained editor.

As illustrated in Fig. 1, we consider a scenario where an image owner relies on a protection service provider to enable authorization-aware use of a publicly released image. In the upper part of the figure, the owner submits an original image to the service, which jointly optimizes a protected image and a key embedding, then maps the key embedding to a user credential, returns the protected image and credential to the owner, and retains the key embedding. At the bottom of the figure, an unauthorized user may directly use the protected image with a DiT editor but cannot obtain reference-consistent edits, whereas an authorized user provides a valid credential to the service, which retrieves the corresponding key embedding and enables reference-consistent editing. Thus, the protected image provides default protection, while the credential enables authorized recovery.

We realize **authorization-aware protection** through **Key-Recoverable Conditional Flow Hijacking**. Our method jointly optimizes a visual perturbation and a key embedding through the native visual- and text-conditioning pathways of a DiT. Because these pathways act through different inputs but jointly influence the same editing process, they can form an asymmetric control pair. This reframes adversarial image protection from one-way disruption to controllable access to reference-consistent editing, without modifying the pretrained editor.

Our main contributions are as follows:

- We formulate **authorization-aware protection for DiT-based in-context editing**, extending adversarial image protection from indiscriminate disruption to selective control of reference-consistent editing.
- We develop **Key-Recoverable Conditional Flow Hijacking**, which exploits the native visual- and text-conditioning pathways of a frozen DiT to achieve unauthorized protection and authorized recovery without modifying or retraining the editor.
- We demonstrate **image-specific authorization selectivity**, showing that recovery is bound to the corresponding image and key: non-matching keys fail to restore reference consistency, while the correct key enables authorized recovery.
- We further show **cross-model authorization transferability**, with both unauthorized protection and authorized recovery transferring to unseen DiT editors, allowing a protection service provider to reuse one image-key pair across models. We also validate robustness to common image transformations and purification.

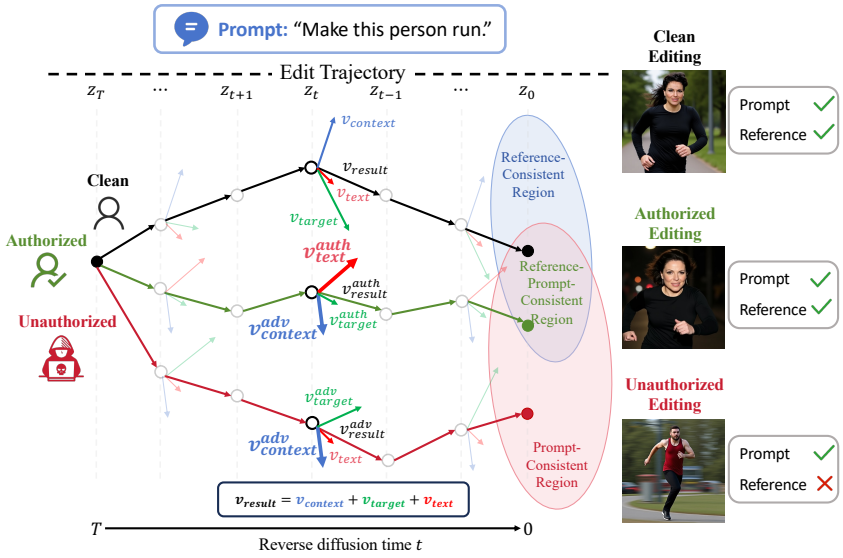


Figure 2: Conceptual editing trajectories under clean, unauthorized, and authorized conditions.

2 Related Work

Existing image protection methods typically add imperceptible perturbations to prevent unauthorized personalization or editing (Van Le et al., 2023; Liu et al., 2024; Salman et al., 2023; Chen et al., 2024), with later work improving robustness or targeting denoising dynamics, identity, frequency-domain protection, and cross-model transfer (Choi et al., 2025; Wang et al., 2025; Hong et al., 2025; Kim et al., 2025; Bala et al., 2025; Shao et al., 2026). Reference preservation can also be disrupted while maintaining prompt following (Tang et al., 2025), and recent DiT-oriented methods such as DeContext and VETO further interfere with reference-context propagation (Shen et al., 2025; Grebe et al., 2026). These methods, however, are largely authorization-agnostic. In contrast, authorization-aware approaches introduce selectivity through semantic or model-side control (Shen et al., 2026; Zeng et al., 2025), invertible or secret-key-based recovery (Hu et al., 2025; Jia et al., 2026), or visual-trigger-based model locking (Le & Fu, 2026). Our method instead binds authorization to each protected reference image using only a visual perturbation and key embedding through the native conditioning pathways of a DiT, without auxiliary models or model-side training.

3 Methodology

3.1 Authorization Formulation

Given a reference image x and a DiT-based in-context editor F_θ, we jointly optimize a visual perturbation δ and a key embedding E_k. The resulting protected image

$$x^{\text{adv}} = x + \delta, \qquad \|\delta\|_\infty \le \epsilon, \tag{1}$$

is intended for public release, while E_k is kept by the protection service for authorized recovery.

Threat model. We assume that an unauthorized user can obtain the publicly released protected image x^{adv}, choose arbitrary editing prompts, and freely use a DiT-based in-context editor, but cannot obtain valid credentials. An authorized user holds the valid credential, which allows the service to retrieve the corresponding key embedding E_k for authorized editing.

Let E_p denote the text embedding of a prompt, and E_k^- denote an absent or non-matching key embedding. Our authorization objective is asymmetric:

$$\begin{aligned} F_\theta(x^{\text{adv}}, [E_p; E_k^-]) &: \quad \text{reference } \textbf{inconsistency} \text{ with prompt consistency,} \\ F_\theta(x^{\text{adv}}, [E_p; E_k]) &: \quad \text{reference consistency with prompt consistency.} \end{aligned} \tag{2}$$

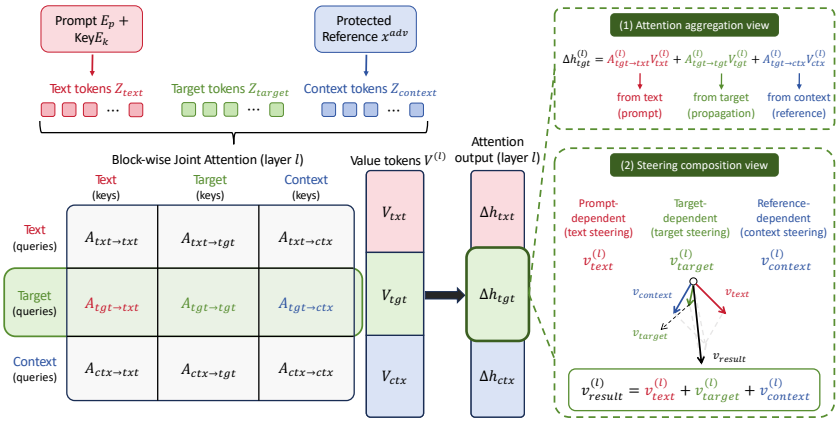


Figure 3: Block-wise attention and target-side steering composition.

Here, E_k^- covers unauthorized conditions such as no key or a non-matching key, whereas $[E_p; E_k]$ denotes appending the correct key embedding to the prompt embedding sequence.

Conceptually, we realize authorization by controlling the editing trajectory induced by the reference condition. As illustrated in Fig. 2, clean editing follows the black trajectory toward the overlap of the reference-(blue region) and prompt-consistent (red region) regions on the right, producing an edit consistent with both reference and prompt. Protection redirects the red unauthorized trajectory away from the reference-consistent region (blue region) while preserving prompt consistency, yielding an edit that follows the prompt but not the reference image. With the correct key, the green authorized trajectory is guided back toward the overlap, restoring consistency with both conditions. This trajectory-level view provides the principle of our protection and recovery mechanism.

3.2 Asymmetric Control through Native Conditioning Pathways

To realize the trajectory-level behavior in Fig. 2, we exploit the native visual- and text-conditioning pathways of a DiT. The protected image and key embedding act through the visual and text pathways, respectively, corresponding to the $v_{\text{context}}^{\text{adv}}$ and $v_{\text{text}}^{\text{auth}}$ steering arrows in the figure. Although entering through distinct inputs, both influence the same target-side update v_{result}, allowing the protected image and key embedding to form an asymmetric control pair. We next examine how these pathways contribute to v_{result} at each transformer block.

As illustrated in Fig. 3, at transformer block ℓ, the hidden sequence is organized into text, target, and context token groups:

$$\mathbf{Z}^{(\ell)} = [\mathbf{Z}_{\text{text}}^{(\ell)}; \mathbf{Z}_{\text{target}}^{(\ell)}; \mathbf{Z}_{\text{context}}^{(\ell)}]. \tag{3}$$

These groups represent the prompt condition, evolving target, and reference context, respectively. Standard joint attention projects them into queries, keys, and values. Let $\mathbf{A}_{i\to j}^{(\ell)}$ denote the attention block from query group i to key group j, and $\mathbf{V}_j^{(\ell)}$ the corresponding value tokens.

The target-query row highlighted in Fig. 3 aggregates information from all three groups and determines the local update of the evolving target. We reinterpret this update as

$$\underbrace{\Delta\mathbf{h}_{\text{target}}^{(\ell)}}_{v_{\text{result}}^{(\ell)}} = \underbrace{\mathbf{A}_{\text{tar}\to\text{txt}}^{(\ell)}\mathbf{V}_{\text{txt}}^{(\ell)}}_{v_{\text{text}}^{(\ell)}} + \underbrace{\mathbf{A}_{\text{tar}\to\text{tar}}^{(\ell)}\mathbf{V}_{\text{tar}}^{(\ell)}}_{v_{\text{target}}^{(\ell)}} + \underbrace{\mathbf{A}_{\text{tar}\to\text{ctx}}^{(\ell)}\mathbf{V}_{\text{ctx}}^{(\ell)}}_{v_{\text{context}}^{(\ell)}}. \tag{4}$$

As visualized in Fig. 3, v_{text}, v_{target}, and v_{context} correspond to prompt-dependent, target-dependent, and reference-dependent steering, respectively, and their sum forms the resulting target-side update v_{result}. Across transformer blocks and denoising steps, these local updates accumulate to form the editing trajectories in Fig. 2. In particular, v_{context} carries reference-dependent information into the evolving target, while v_{text} provides a separate route through which the key embedding can influence the same target-side update. Steering here denotes a local representation update rather than a direct decomposition of the model's flow velocity; its functional roles are validated in Sec. 4.

This decomposition connects the trajectory-level behavior in Fig. 2 with the block-level control in Fig. 3. We redirect the reference-dependent v_{context} to drive unauthorized editing away from clean-reference behavior, while the key acts through text conditioning to adjust v_{result} for authorized recovery. These two effects motivate the protection and recovery objectives introduced next.

3.3 Unauthorized Protection via Conditional Flow Hijacking

We first construct the unauthorized branch by hijacking the reference-induced context steering v_{context}. Instead of suppressing context interaction, we preserve context engagement while redirecting its steering away from clean-reference behavior, thereby disrupting reference-consistent editing without the correct key.

We introduce a hijack anchor x^{hijack}, instantiated as an unrelated image or random noise, and extract its context steering $v_{\text{context}}^{\text{hijack}}$ under the same editing condition to provide a steering-space reference rather than a target output. For the protected image x^{adv}, we define

$$\mathcal{L}_{\text{hij}} = \frac{1}{|\mathcal{S}|} \sum_{\ell \in \mathcal{S}} \text{MSE}\left(v_{\text{context}}^{\text{adv},(\ell)}, v_{\text{context}}^{\text{hijack},(\ell)}\right), \tag{5}$$

where $\mathcal{S}$ denotes the sampled transformer blocks. To further separate the protected steering from clean-reference behavior, we define

$$\mathcal{L}_{\text{rep}} = \frac{1}{|\mathcal{S}|} \sum_{\ell \in \mathcal{S}} \text{MSE}\left(v_{\text{context}}^{\text{adv},(\ell)}, v_{\text{context}}^{\text{clean},(\ell)}\right). \tag{6}$$

To prevent the optimization from collapsing into trivial context suppression, we preserve target-to-context interaction, so protection redirects context steering rather than removing context engagement. Following DeContext (Shen et al., 2025), we measure context engagement as

$$r_{\text{ctx}} = \frac{1}{H|\mathcal{S}||\mathcal{Q}|} \sum_{\ell \in \mathcal{S}} \sum_{m=1}^{H} \sum_{q \in \mathcal{Q}} \sum_{k \in \mathcal{C}} A_{tgt \to ctx}^{(\ell,m)}, \tag{7}$$

where H is the number of attention heads, and $\mathcal{Q}$ and $\mathcal{C}$ denote the target-query and context-key sets. Unlike DeContext, which reduces r_{ctx}, we preserve this pathway with

$$\mathcal{L}_{\text{attn}} = 1 - r_{\text{ctx}}. \tag{8}$$

The overall unauthorized protection objective is

$$\mathcal{L}_{\text{prot}} = \lambda_{\text{attn}} \mathcal{L}_{\text{attn}} + \lambda_{\text{hij}} \mathcal{L}_{\text{hij}} - \lambda_{\text{rep}} \mathcal{L}_{\text{rep}}. \tag{9}$$

Minimizing $\mathcal{L}_{\text{prot}}$ preserves context engagement while steering the protected context dynamics toward the hijack anchor and away from the clean reference. We optimize δ with projected sign-gradient updates under the ℓ_∞ constraint (Madry et al., 2018), establishing the default unauthorized state before key recovery.

3.4 Authorized Recovery via Key Conditioning

We next construct the authorized branch of asymmetric control. For each protected image, we optimize a key embedding $E_k \in \mathbb{R}^{L_k \times d_t}$ through the model's native text-conditioning pathway, where L_k and d_t denote the key length and text-embedding dimension, respectively. The key is learned to re-coordinate the target-side dynamics disrupted by the protected visual condition.

Given the prompt embedding E_p, authorized editing appends the key as

$$E_p^{\text{key}} = [E_p; E_k], \tag{10}$$

so the key enters the same joint attention pathway as the editing prompt.

To guide recovery, we use the clean reference under the original prompt as a local dynamical reference. For each sampled block $\ell \in \mathcal{S}$, let $v_{\text{result}}^{\text{clean},(\ell)}$ denote the target-side update from the clean image and E_p, and $v_{\text{result}}^{\text{auth},(\ell)}$ that from x^{adv} and E_p^{key}. We optimize

$$\mathcal{L}_{\text{auth}} = \frac{1}{|\mathcal{S}|} \sum_{\ell \in \mathcal{S}} \text{MSE}\left(v_{\text{result}}^{\text{auth},(\ell)}, v_{\text{result}}^{\text{clean},(\ell)}\right). \tag{11}$$

Algorithm 1 Alternating Image-Key Optimization

Require: Clean image x, hijack anchor x^{hijack}, prompt pool $\mathcal{P}$, timestep set $\mathcal{T}$, frozen editor F_θ, perturbation budget ϵ

Initialize $x^{\text{adv}} \leftarrow x$ and randomly initialize the key E_k
for $s = 1, \ldots, N$ **do**
  Sample $p \sim \mathcal{P}$, $t \sim \mathcal{T}$, and blocks $\mathcal{S}$
  Sample a non-matching key $E_k^- \sim \mathcal{D}_{E_k}^-$ and set $E_p^- = [E_p; E_k^-]$
  Extract $v_{\text{context}}^{\text{hijack}}$, $v_{\text{context}}^{\text{clean}}$, and $v_{\text{context}}^{\text{adv}}$ under E_p^-
  Compute r_{ctx} and $\mathcal{L}_{\text{prot}}$ using Eqs. 7 and 9
  $\widetilde{x}^{\text{adv}} \leftarrow x^{\text{adv}} - \alpha \ \text{sign}\left(\nabla_{x^{\text{adv}}} \mathcal{L}_{\text{prot}}\right)$
  $x^{\text{adv}} \leftarrow \text{clip}_{[-1,1]}\left(x + \text{clip}_{[-\epsilon,\epsilon]}\left(\widetilde{x}^{\text{adv}} - x\right)\right)$
  Set $E_p^{\text{key}} = [E_p; E_k]$
  Extract $v_{\text{result}}^{\text{clean}}$ from (x, E_p) and $v_{\text{result}}^{\text{auth}}$ from $(x^{\text{adv}}, E_p^{\text{key}})$
  Compute $\mathcal{L}_{\text{auth}}$ using Eq. 11
  Update E_k using $\nabla_{E_k}\left(\lambda_{\text{auth}} \mathcal{L}_{\text{auth}}\right)$
end for
return Protected image x^{adv} and key E_k

Minimizing $\mathcal{L}_{\text{auth}}$ locally re-aligns the protected target-side dynamics with clean-reference behavior. The clean dynamics serve as a recovery reference rather than requiring exact trajectory reproduction. Together with Sec. 3.3, this yields image-specific authorization within a fully frozen editor without introducing additional networks. We further analyze the resulting protection and recovery behavior in Appendix A.1, and provide component-wise ablations of the loss design in A.2.

3.5 Alternating Image-Key Optimization

We construct the protected image and its key through alternating optimization. Let $\mathcal{P}$ denote a prompt pool and $\mathcal{T}$ a set of high-noise timesteps from the early denoising stage. Following prior observations that early denoising steps are particularly influential for adversarial protection and context propagation (Choi et al., 2025), at each iteration we sample $p \sim \mathcal{P}$, $t \sim \mathcal{T}$, and a subset of transformer blocks $\mathcal{S}$, and share these conditions across the corresponding steering comparisons.

For the protected-image update, we additionally sample a non-matching key $E_k^- \sim \mathcal{D}_{E_k}^-$ and form the unauthorized condition $E_p^- = [E_p; E_k^-]$, where $\mathcal{D}_{E_k}^-$ denotes the sampling distribution of non-matching keys with the same shape as the key. Resampling E_k^- across iterations encourages protection to remain effective under diverse mismatched-key conditions.

The alternating construction is formulated as

$$\begin{aligned} \delta^* &\in \arg\min_{\|\delta\|_\infty \le \epsilon} \mathbb{E}_{p\sim\mathcal{P},\, t\sim\mathcal{T},\, \mathcal{S},\, E_k^- \sim \mathcal{D}_{E_k}^-} \left[\mathcal{L}_{\text{prot}}(\delta; E_k^-)\right], \\ E_k^* &\in \arg\min_{E_k} \mathbb{E}_{p\sim\mathcal{P},\, t\sim\mathcal{T},\, \mathcal{S}} \left[\lambda_{\text{auth}} \mathcal{L}_{\text{auth}}(\delta^*, E_k)\right]. \end{aligned} \tag{12}$$

Here, $\mathcal{L}_{\text{prot}}$ constructs the default unauthorized protection, while $\mathcal{L}_{\text{auth}}$ optimizes the key for authorized recovery on the resulting protected image. We solve the two objectives alternately, as summarized in Alg. 1. Sampling over prompts, timesteps, transformer blocks, and non-matching keys reduces dependence on particular editing or unauthorized-key conditions.

4 Functional Disentanglement of Conditional Steering

Context and text steering play functionally disentangled roles in in-context editing: context steering primarily governs reference-dependent conditioning, while text steering primarily governs prompt-dependent editing. This functional separation provides the basis for our asymmetric control through the visual and text-conditioning pathways. We verify this property through *counterfactual steering transplantation* during clean editing.

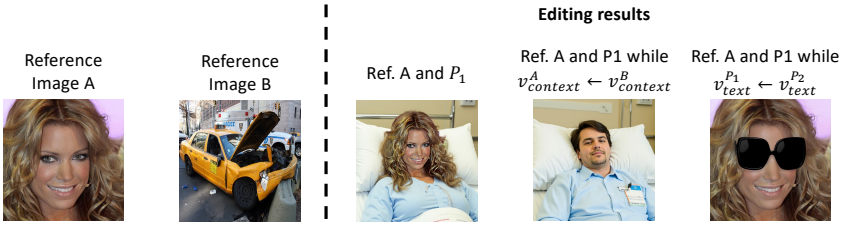


Figure 4: Counterfactual steering transplantation. Editing prompts P_1: "a person in the hospital." and P_2: "let the person wear sunglasses."

Let A and B denote two reference images, and P_1 and P_2 two editing prompts. For context steering, we edit both references under P_1 and replace the context steering of Reference A with v^B_{context}. For text steering, we keep Reference A fixed and replace its text steering under P_1 with $v^{P_2}_{\text{text}}$. Inference uses 20 denoising steps, with transplantation applied during the first 5 steps to 25 randomly sampled transformer blocks per step.

As shown in Fig. 4, transplanting v^B_{context} substantially changes the generated subject while largely preserving the semantics specified by P_1. In contrast, transplanting $v^{P_2}_{\text{text}}$ introduces the semantics of P_2 while largely preserving the reference subject. These results support the functional disentanglement of context and text steering and justify their roles in the asymmetric control pair. Additional analyses of context-steering intervention, hijacking, and key-conditioned recovery are provided in Appendix A.1.

5 Experiments

5.1 Experimental Setup

Datasets and prompts. We evaluate on CelebA-HQ (Liu et al., 2015) and VGGFace2 (Cao et al., 2018), sampling 50 reference images per dataset at 512×512. We use two standard portrait prompts, "a photo of this person" and "a dslr portrait of this person", together with 60 held-out prompts unseen during optimization, spanning attribute, pose/action, object interaction, appearance, scene, and combined edits. Each prompt is applied to all references.

Models and baselines. We conduct the main evaluation on FLUX.1-Kontext and compare against AdvDM, EditShield, FaceLock, and DeContext. We further evaluate cross-model transferability on unseen DiT-based editors (Mao et al., 2025; Xia et al., 2026; Zhang et al., 2025; Liu et al., 2025).

Implementation details. We use perturbation budget $\epsilon = 0.1$, PGD step size $\alpha = 0.005$, and 200 optimization iterations, sampling $t \in \{980, \ldots, 1000\}$ and 25 transformer blocks per iteration. We set $\lambda_{\text{attn}} = 0.01$ and $\lambda_{\text{hij}} = \lambda_{\text{rep}} = \lambda_{\text{auth}} = 1$. Each image is paired with an independently optimized key embedding of length 16 and dimension 4096.

Evaluation metrics. We report Face Detection Failure Rate (FDFR), Identity Score Matching (ISM), CLIP-IQA (Wang et al., 2023), and CLIP-Score (Hessel et al., 2021). FDFR is measured with RetinaFace (Deng et al., 2020); for samples with a detected face, ISM is computed using ArcFace (Deng et al., 2019) to measure identity similarity to the reference. CLIP-IQA and CLIP-Score measure perceptual quality and prompt alignment, respectively. Under unauthorized editing, stronger protection corresponds to lower ISM while maintaining quality and prompt alignment; under authorized editing, successful recovery should move ISM toward the clean result while maintaining low FDFR and competitive CLIP-IQA and CLIP-Score.

5.2 Authorization-Aware Protection and Recovery

Across standard and diverse prompts, the same protected image exhibits the intended authorization-dependent behavior: unauthorized editing disrupts reference identity preservation, while the Correct Key restores it toward the Clean result.

Table 1: Quantitative comparison under standard and diverse editing prompts.

(a) Quantitative results under standard portrait prompts.

Dataset	Method	"a photo of this person"				"a dslr portrait of this person"			
		FDFR	ISM↓	CLIP-IQA↑	CLIP-Score↑	FDFR	ISM↓	CLIP-IQA↑	CLIP-Score↑
CelebA-HQ	Clean	0.00	0.89	0.78	24.13	0.00	0.64	0.81	23.96
	AdvDM	0.00	0.87	0.59	24.42	0.00	0.60	0.80	24.10
	EditShield	0.00	0.93	**0.72**	24.20	0.00	0.60	**0.81**	24.00
	FaceLock	0.00	0.86	0.70	24.27	0.00	0.67	0.73	22.21
	DeContext	0.00	0.39	0.70	25.93	0.00	0.41	**0.81**	24.85
	Ours (No Key)	0.00	**0.14**	0.71	**26.57**	0.00	**0.18**	0.78	**25.40**
	Ours (Correct Key)	0.00	0.77	0.70	24.37	0.00	0.57	0.77	24.15
VGGFace2	Clean	0.00	0.97	0.79	23.22	0.00	0.82	0.80	22.48
	AdvDM	0.00	0.90	0.65	24.26	0.00	0.75	**0.80**	22.80
	EditShield	0.00	0.94	**0.74**	23.74	0.00	0.79	0.78	22.85
	FaceLock	0.00	0.90	0.64	24.09	0.00	0.89	0.68	22.01
	DeContext	0.00	0.52	0.70	25.12	0.00	0.50	0.77	23.72
	Ours (No Key)	0.00	**0.17**	0.69	**26.33**	0.00	**0.24**	0.71	**24.64**
	Ours (Correct Key)	0.00	0.83	0.68	23.71	0.00	0.72	0.77	22.75

(b) Quantitative results under diverse held-out editing prompts.

Method	CelebA-HQ				VGGFace2			
	FDFR	ISM↓	CLIP-IQA↑	CLIP-Score↑	FDFR	ISM↓	CLIP-IQA↑	CLIP-Score↑
Clean	0.00	0.78	0.81	26.56	0.00	0.79	0.82	25.65
AdvDM	0.00	0.71	0.72	26.75	0.01	0.72	0.77	26.10
EditShield	0.00	0.76	**0.78**	26.67	0.00	0.77	**0.82**	25.91
FaceLock	0.00	0.72	0.75	26.54	0.01	0.71	0.75	25.97
DeContext	0.01	0.57	0.76	27.22	0.03	0.56	0.78	26.80
Ours (No Key)	0.09	**0.26**	0.76	**27.67**	0.09	**0.34**	0.75	**27.29**
Ours (Correct Key)	0.00	0.68	0.70	26.46	0.00	0.70	0.73	25.74

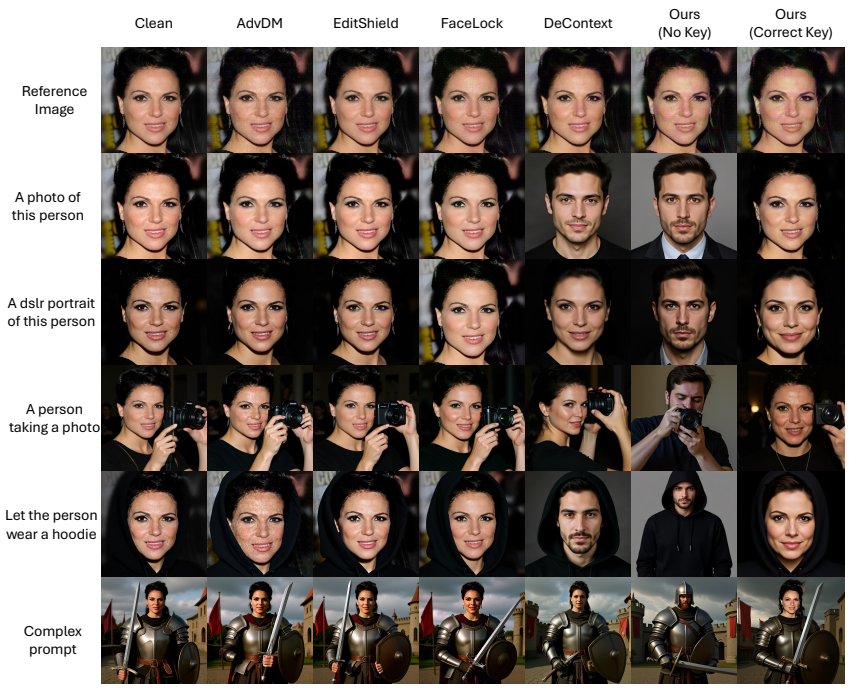


Figure 5: Qualitative comparison under standard and diverse editing prompts. The final row uses the complex prompt: "Dress this person as a medieval knight with sword and shield, standing in a castle courtyard with banners and dramatic sky."

As shown in Tabs. 1a and 1b, Ours exhibits consistent authorization-dependent behavior across standard and diverse prompts. Under the standard "photo" prompt, Ours (No Key) reduces ISM from 0.89/0.97 to 0.14/0.17 on CelebA-HQ/VGGFace2, below DeContext (0.39/0.52), while the Correct Key recovers it to 0.77/0.83. Under diverse held-out edits, ISM similarly drops from 0.78/0.79 to 0.26/0.34 and recovers to 0.68/0.70 with the Correct Key, showing that the same protected image blocks reference-consistent editing by default while remaining recoverable across prompt regimes.

Table 2: Authorization selectivity under two standard prompts on CelebA-HQ.

Condition	"a photo of this person"				"a dslr portrait of this person"			
	FDFR	ISM↑	CLIP-IQA↑	CLIP-Score↑	FDFR	ISM↑	CLIP-IQA↑	CLIP-Score↑
Clean	0.00	0.89	0.78	24.13	0.00	0.64	0.81	23.96
No Key	0.00	0.14	0.71	**26.57**	0.00	0.18	**0.78**	**25.40**
Random Key	0.00	0.14	0.66	25.74	0.00	0.18	0.76	25.16
Swap Key	0.00	0.20	**0.72**	25.22	0.00	0.23	0.75	24.12
Correct Key	0.00	**0.77**	0.70	24.37	0.00	**0.57**	0.77	24.15

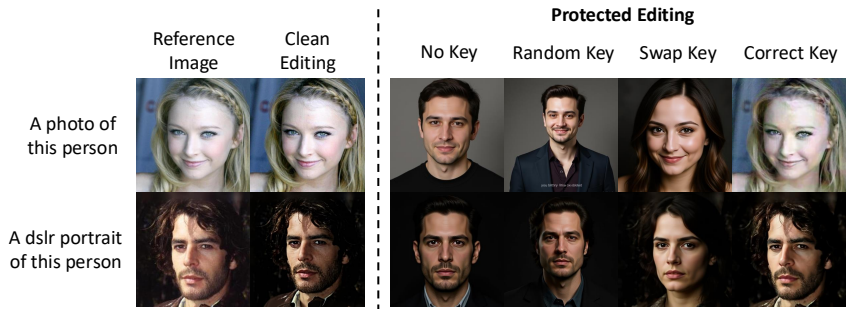


Figure 6: Qualitative analysis of image-specific authorization selectivity.

Fig. 5 shows consistent qualitative behavior: across standard, diverse, and complex edits, Ours (No Key) follows the prompt while producing a reference-inconsistent identity, whereas the Correct Key restores reference-consistent appearance under the same prompt.

5.3 Image-Specific Authorization Selectivity

Authorized recovery is specific to the corresponding image-key pair rather than arbitrary key conditioning. We evaluate this selectivity on CelebA-HQ under two standard prompts, comparing No Key, Random Key, Swap Key, and Correct Key conditions, with Clean as reference. Random Key denotes a sampled non-matching key, while Swap Key uses the key of another protected image.

As shown in Tab. 2, non-matching keys consistently fail to restore reference identity. Under the "photo" prompt, No Key, Random Key, and Swap Key yield ISM values of 0.14, 0.14, and 0.20, whereas the Correct Key recovers to 0.77 toward the Clean level of 0.89. The same pattern holds for the DSLR prompt (0.18/0.18/0.23 versus 0.57, with Clean at 0.64). FDFR remains 0.00 throughout with competitive CLIP-IQA and CLIP-Score, indicating that failed authorization does not simply prevent valid or prompt-aligned generation.

Fig. 6 provides consistent qualitative evidence: No Key, Random Key, and Swap Key all produce plausible edited faces but with identities inconsistent with the protected reference, whereas only the Correct Key restores reference-consistent identity. This confirms that authorized recovery is bound to the specific image-key pair rather than generic key conditioning.

5.4 Additional Analyses

We further validate cross-model authorization transferability on four unseen DiT-based editors. On ICEdit, ISM drops from 0.91 to 0.49 under No Key and recovers to 0.77 with the same FLUX.1-Kontext key, without any target-model re-optimization. We also study sensitivity to key design and robustness to image transformations and purification, with full results in Appendix A.

6 Conclusion

We present **Key-Recoverable Conditional Flow Hijacking** for authorization-aware DiT-based in-context editing. Our method enables unauthorized protection and key-specific authorized recovery through the native conditioning pathways of a frozen DiT. Experiments verify effective protection, selective recovery, cross-model transferability and robustness.

AI USE STATEMENT

In this work, we **used** generative AI tools to help develop and refine conceptual frameworks, propose and refine research hypotheses, design and provide feedback on research methodology and experiments, assist with implementing parts of the proposed method, assist with translation, support qualitative analysis, and interpret experimental results. We have **not used** generative AI tools to generate synthetic datasets, formulate mathematical claims, or clean and reformat datasets, and tasks involving critical ingredients for proving mathematical claims and the writing of proofs are not applicable to this work.

Additionally, we **used** generative AI tools to create and edit code, draft and edit parts of the paper, summarize and analyze existing literature, brainstorm research ideas, source and search for information, improve the readability of the manuscript, identify relevant literature, and suggest the structure of the paper. We have reviewed all AI-assisted work. AI-assisted code was manually reviewed and tested by the authors; research ideas, methodological suggestions, literature references, and result interpretations were also manually checked and verified, and all reported experimental results were obtained and validated by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the methodological details and optimization procedure in Sec. 3, including the complete alternating image-key optimization in Alg. 1. Experimental settings, including datasets, evaluation prompts, implementation settings, and evaluation metrics, are described in Sec. 5.1. Additional analyses, including mechanism analysis, ablations, sensitivity studies, robustness tests, and cross-model transferability, are provided in Appendix A, together with the complete set of evaluation prompts. We will provide the core implementation code and scripts needed to reproduce the main experiments in the supplementary material.

REFERENCES

Aniruddha Bala, Rohit Chowdhury, Rohan Jaiswal, and Siddharth Roheda. DCT-Shield: a robust frequency domain defense against malicious image editing. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18876–18884. IEEE, 2025.

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.

Bochuan Cao, Changjiang Li, Ting Wang, Jinyuan Jia, Bo Li, and Jinghui Chen. IMPRESS: Evaluating the resilience of imperceptible perturbations against unauthorized data usage in diffusion-based generative AI. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 10657–10677. Curran Associates, Inc., 2023. doi: 10.52202/075280-0468. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/222dda29587fbc2979ca99fd5ed00735-Paper-Conference.pdf.

Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 67–74, 2018. doi: 10.1109/FG.2018.00020.

Ruoxi Chen, Haibo Jin, Yixin Liu, Jinyin Chen, Haohan Wang, and Lichao Sun. EditShield: Protecting unauthorized image editing by instruction-guided diffusion models. In *European Conference on Computer Vision*, pp. 126–142. Springer, 2024.

William June Suk Choi, Kyungmin Lee, Jongheon Jeong, Saining Xie, Jinwoo Shin, and Kimin Lee. DiffusionGuard: A robust defense against malicious diffusion-based image editing. In *International Conference on Learning Representations*, volume 2025, pp. 27134–27180, 2025.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4685–4694, 2019. doi: 10.1109/CVPR.2019.00482.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-shot multi-level face localisation in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5202–5211, 2020. doi: 10.1109/CVPR42600.2020.00525.

Jonas Grebe, Hossein Shakibania, Tobias Braun, Marcus Rohrbach, and Anna Rohrbach. VETO: Towards protecting images from frontier AI editing. *arXiv preprint arXiv:2607.27292*, 2026.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

Seunghoo Hong, Geonho Son, Juhun Lee, and Simon S Woo. DIA: The adversarial exposure of deterministic inversion in diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17994–18003. IEEE, 2025.

Baoyue Hu, Yang Wei, Junhao Xiao, Wendong Huang, Xiuli Bi, and Bin Xiao. Who controls the authorization? invertible networks for copyright protection in text-to-image synthesis. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15832–15841. IEEE, 2025.

Jun Jia, Hongyi Miao, Yingjie Zhou, Wangqiu Zhou, Jianbo Zhang, Linhan Cao, Dandan Zhu, Hua Yang, Xiongkuo Min, Wei Sun, et al. Adapter Shield: A unified framework with built-in authentication for preventing unauthorized zero-shot image-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30120–30129, 2026.

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

Van Thanh Le and Yun Fu. VisiLock: Authorizing instruction-based image editing with dual score distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15710–15718, 2026.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1X-Edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24219–24228. IEEE, 2024.

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

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

Chaojie Mao, Jingfeng Zhang, Yulin Pan, Zeyinzi Jiang, Zhen Han, Yu Liu, and Jingren Zhou. ACE++: Instruction-based image creation and editing via context-aware content filling. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1979–1987. IEEE, 2025.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16805–16827. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/nie22a.html.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious AI-powered image editing. In *The 40th International Conference on Machine Learning (ICML)*. JMLR.org, 2023.

Mingwen Shao, Lingzhuang Meng, Xiang Lv, Mengyao Wu, Xinyuan Chen, Qiao Zhang, Chang Liu, Yuanjian Qiao, and Chao Dong. UniDef: Universal defense against unauthorized image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8631–8640, 2026.

Kaixin Shen, Ruijie Quan, Jiaxu Miao, and Jun Xiao. TarPro: Targeted protection against malicious image editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8896–8904, 2026.

Linghui Shen, Mingyue Cui, and Xingyi Yang. DeContext as defense: Safe image editing in diffusion transformers. *arXiv preprint arXiv:2512.16625*, 2025.

Qiuyu Tang, Bonor Ayambem, Mooi Choo Chuah, and Aparna Bharati. Is perturbation-based image protection disruptive to image editing? In *2025 IEEE International Conference on Image Processing (ICIP)*, pp. 2073–2078. IEEE, 2025.

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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2116–2127. IEEE, 2023.

Hanhui Wang, Yihua Zhang, Ruizheng Bai, Yue Zhao, Sijia Liu, and Zhengzhong Tu. Edit away and my face will not stay: Personal biometric defense against malicious generative editing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23806–23816. IEEE, 2025.

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

Bin Xia, Bohao Peng, Yuechen Zhang, Junjia Huang, Jiyang Liu, Jingyao Li, Haoru Tan, Sitong Wu, Chengyao Wang, Yitong Wang, Bei Yu, and Jiaya Jia. DreamOmni2: Multimodal instruction-based generation and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29275–29284, June 2026.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. OmniGen: Unified image generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13294–13304. IEEE, 2025.

Yaopei Zeng, Yuanpu Cao, and Lu Lin. GuardDoor: Safeguarding against malicious diffusion editing via protective backdoors. *arXiv preprint arXiv:2503.03944*, 2025.

Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-Context Edit: Enabling instructional image editing with in-context generation in large-scale diffusion transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2504.20690.

Zhengyue Zhao, Jinhao Duan, Kaidi Xu, Chenan Wang, Rui Zhang, Zidong Du, Qi Guo, and Xing Hu. Can protective perturbation safeguard personal data from being exploited by Stable Diffusion? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24398–24407, June 2024.

A APPENDIX

Contents

A.1 Mechanism Analysis **14**

A.1.1 Context-Steering Intervention 14

A.1.2 Steering Hijacking and Key-Conditioned Recovery 14

A.2 Ablation Study **15**

A.3 Cross-Model Authorization Transferability **15**

A.4 Sensitivity Analysis **16**

A.4.1 Sensitivity to the Perturbation Budget 16

A.4.2 Sensitivity to Key Length 17

A.4.3 Sensitivity to the Attacked Denoising Stage 17

A.4.4 Sensitivity to the Number of Optimization Steps 17

A.4.5 Sensitivity to the Number of Sampled Blocks 18

A.5 Robustness to Image Transformations and Purification **18**

A.6 Additional Visual Results **19**

A.7 Evaluation Prompts **20**

A.1 Mechanism Analysis

Building on the functional disentanglement analysis in Sec. 4, we further examine whether the altered context steering directly mediates the protection behavior, and whether the proposed objectives produce the intended hijacking and recovery effects.

A.1.1 Context-Steering Intervention

We further quantify the role of context steering by exchanging it between clean and protected branches. For each reference image, we construct two interventions under the same prompt. First, we perform clean editing while replacing its early v_{context} with the corresponding steering extracted from the protected image. Second, we edit the protected image while replacing its early v_{context} with the corresponding steering from the clean image. In both cases, replacement is applied during the first 5 of 20 denoising steps over 25 sampled transformer blocks.

Table 3: Context-steering replacement between clean and protected branches.

Condition	FDFR	ISM	CLIP-IQA	CLIP-Score
Clean + Protected v_{context}	0.00	0.14	0.76	26.88
Protected + Clean v_{context}	0.00	0.78	0.76	24.58

The two replacements produce opposite effects. Replacing clean context steering with protected steering reduces ISM to 0.14, matching the unauthorized protected result, while maintaining image quality and prompt alignment. Conversely, replacing protected context steering with clean steering raises ISM to 0.78, close to the authorized result obtained with the Correct Key, even though the protected image itself remains unchanged. These results show that the protection behavior is largely mediated by the altered reference-dependent context steering: protected steering induces reference inconsistency, while clean steering restores reference-consistent editing.

A.1.2 Steering Hijacking and Key-Conditioned Recovery

We next verify that the proposed objectives produce the intended unauthorized hijacking and authorized recovery effects over the denoising trajectory. Our asymmetric control preserves context interaction while redirecting context steering under unauthorized editing, and uses the key to recover target-side dynamics for authorized editing.

We first measure context engagement using the target-to-context attention proportion r_{ctx} in Eq. 7, comparing Clean, DeContext, and Ours. We quantify steering hijacking by

$$d^{\text{ctx}}_{\text{clean}} = \text{MSE}\left(v^{\text{adv}}_{\text{context}}, v^{\text{clean}}_{\text{context}}\right), \qquad d^{\text{ctx}}_{\text{hijack}} = \text{MSE}\left(v^{\text{adv}}_{\text{context}}, v^{\text{hijack}}_{\text{context}}\right), \tag{13}$$

and evaluate recovery by the MSE between the protected v_{result} and $v^{\text{clean}}_{\text{result}}$ under No Key, Wrong Key, and Correct Key. All quantities are averaged over trajectories and examples.

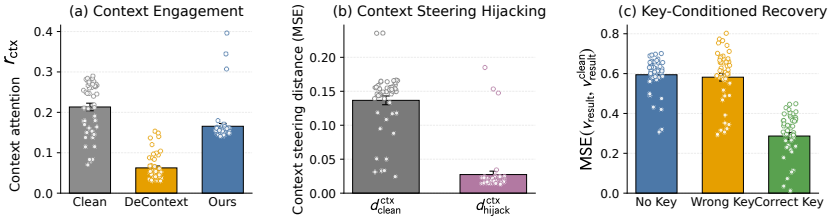


Figure 7: Analysis of context steering hijacking and key-conditioned recovery.

As shown in Fig. 7, Ours preserves stronger context engagement than DeContext, while $d^{\text{ctx}}_{\text{clean}} > d^{\text{ctx}}_{\text{hijack}}$ indicates that protected context steering is redirected toward the hijack anchor. Moreover, No Key and Wrong Key remain far from $v^{\text{clean}}_{\text{result}}$, whereas the Correct Key substantially reduces this discrepancy. These results validate unauthorized steering hijacking and key recovery as complementary mechanisms of asymmetric authorization.

A.2 Ablation Study

To assess the contribution of each component in the unauthorized protection objective, we individually remove $\mathcal{L}_{\mathrm{attn}}$, $\mathcal{L}_{\mathrm{rep}}$, and $\mathcal{L}_{\mathrm{hij}}$ while keeping the remaining settings unchanged. We conduct the study on CelebA-HQ using the diverse held-out editing prompts, and evaluate both unauthorized protection under No Key and authorized recovery under the Correct Key.

Table 4: Ablation study.

Method	No Key				Correct Key			
	FDFR	ISM↓	CLIP-IQA↑	CLIP-Score↑	FDFR	ISM↑	CLIP-IQA↑	CLIP-Score↑
Clean	0.00	0.78	0.81	26.56	–	–	–	–
w/o $\mathcal{L}_{\mathrm{attn}}$	0.10	0.42	0.72	27.50	0.00	0.68	0.69	26.43
w/o $\mathcal{L}_{\mathrm{rep}}$	0.03	0.47	0.75	27.46	0.00	**0.73**	0.69	26.42
w/o $\mathcal{L}_{\mathrm{hij}}$	0.01	0.73	0.71	26.35	0.00	0.68	**0.75**	**26.84**
Ours	0.09	**0.26**	**0.76**	**27.67**	0.00	0.68	0.70	26.46

Overall, removing any protection term weakens unauthorized identity protection, while Correct Key recovery remains relatively stable. As shown in Tab. 4, $\mathcal{L}_{\mathrm{hij}}$ is the most critical component: removing it increases No Key ISM from 0.26 to 0.73, close to the Clean level of 0.78, showing that attraction toward hijack steering is the primary driver of reference disruption. Removing $\mathcal{L}_{\mathrm{rep}}$ increases ISM to 0.47, indicating that explicitly separating protected steering from clean-reference behavior further strengthens protection. Removing $\mathcal{L}_{\mathrm{attn}}$ increases ISM to 0.42, confirming that preserving target-to-context interaction through r_{ctx} facilitates effective steering redirection while maintaining context engagement. Meanwhile, Correct Key ISM stays within 0.68 to 0.73 across all variants, suggesting that these objectives mainly control unauthorized protection strength without substantially impairing authorized recovery. The full objective therefore yields the strongest authorization separation, achieving the lowest No Key ISM of 0.26 while recovering to 0.68 with the Correct Key.

A.3 Cross-Model Authorization Transferability

Our image-key pair exhibits clear cross-model transferability. Although both the protected image and key embedding are optimized only on FLUX.1-Kontext, the protected image remains effective without the key on unseen editors, while the same key consistently restores reference identity. This allows a protection service provider to reuse one key embedding across multiple compatible editors without model-specific key optimization.

Table 5: Cross-model transferability of the protected image and key embedding optimized on FLUX.1-Kontext.

Model	Setting	No Key				Correct Key			
		FDFR	ISM↓	CLIP-IQA↑	CLIP-Score↑	FDFR	ISM↑	CLIP-IQA↑	CLIP-Score↑
FLUX.1-Kontext	Clean	0.00	0.89	0.78	24.13	–	–	–	–
	Protected	0.00	0.14	0.71	26.57	0.00	0.77	0.70	24.37
ACE++	Clean	0.04	0.74	0.76	22.92	–	–	–	–
	Protected	0.02	0.42	0.63	22.85	0.00	0.64	0.70	22.97
DreamOmni2	Clean	0.00	0.91	0.77	23.88	–	–	–	–
	Protected	0.14	0.45	0.66	24.14	0.22	0.75	0.67	24.29
ICEdit	Clean	0.00	0.91	0.78	23.46	–	–	–	–
	Protected	0.00	0.49	0.49	23.57	0.00	0.77	0.59	23.31
Step1X-Edit	Clean	0.00	0.83	0.75	24.36	–	–	–	–
	Protected	0.00	0.42	0.54	23.73	0.00	0.63	0.61	23.43

As shown in Tab. 5, the protected image substantially reduces identity similarity across all evaluated unseen editors. Compared with clean editing, the No Key ISM decreases from 0.74 to 0.42 on ACE++, from 0.91 to 0.45 on DreamOmni2, from 0.91 to 0.49 on ICEdit, and from 0.83 to 0.42 on Step1X-Edit. Importantly, the same key embedding optimized on FLUX.1-Kontext restores ISM to 0.64, 0.75, 0.77, and 0.63, respectively, without any target-editor key optimization. CLIP-IQA and CLIP-Score remain broadly comparable across the No Key and Correct Key conditions, indicating that recovery does not rely on sacrificing perceptual quality or prompt alignment.

Overall, these results show that both unauthorized protection and authorized recovery transfer beyond the source editor. In particular, the portability of the same key embedding allows a protection service provider to support multiple compatible DiT-based editors with a single image-key pair, providing greater flexibility for multi-editor deployment.

A.4 Sensitivity Analysis

We further analyze the sensitivity of our method to several key design and optimization choices, including the perturbation budget ϵ, the number of sampled transformer blocks, the attacked denoising stage, the number of perturbation optimization iterations, and the key length. These analyses characterize how protection strength, authorized recovery, and image fidelity vary under different configurations.

A.4.1 Sensitivity to the Perturbation Budget

We first study the effect of the perturbation budget ϵ, which directly controls the allowable deviation from the clean reference and therefore governs the trade-off between protection strength and image fidelity. We evaluate $\epsilon \in \{0.02, 0.04, 0.06, 0.08, 0.10\}$ on CelebA-HQ under the standard "a photo of this person" prompt, measuring both protected-image fidelity and editing performance under No Key and Correct Key conditions.

Table 6: Sensitivity to the perturbation budget ϵ.

ϵ	PSNR↑	SSIM↑	LPIPS↓	No Key		Correct Key	
				ISM↓	CLIP-IQA↑	ISM↑	CLIP-IQA↑
0.00 (Clean)	–	–	–	0.89	0.78	–	–
0.02	**42.39**	**0.99**	**0.16**	0.71	**0.75**	**0.97**	**0.75**
0.04	36.50	0.96	0.33	0.42	0.72	0.91	0.72
0.06	33.35	0.92	0.42	0.22	0.71	0.80	0.71
0.08	31.13	0.88	0.48	**0.12**	0.72	0.80	0.70
0.10	29.54	0.84	0.51	0.14	0.71	0.77	0.70

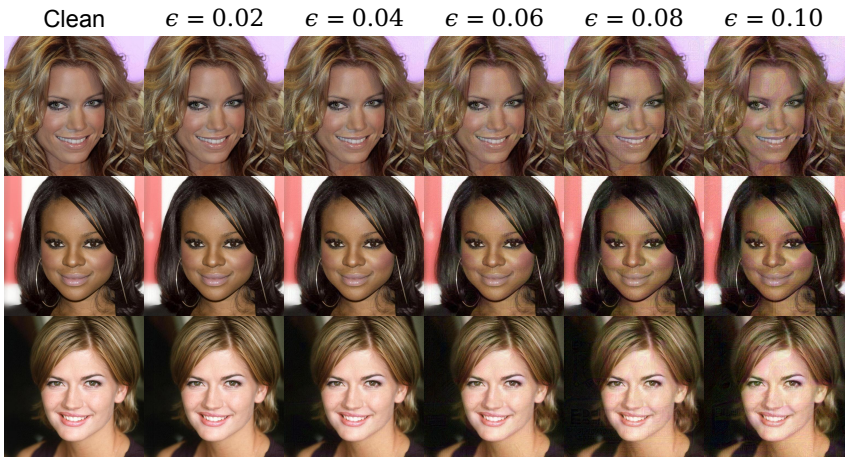


Figure 8: Protected images under different perturbation budgets.

Overall, increasing ϵ substantially strengthens unauthorized protection while preserving effective Correct Key recovery, at the cost of gradually reduced protected-image fidelity. As shown in Tab. 6, the No Key ISM decreases from 0.71 at $\epsilon = 0.02$ to 0.22 at $\epsilon = 0.06$, while the Correct Key ISM remains relatively high at 0.80. At $\epsilon = 0.06$, the protected images still retain relatively high fidelity, with a PSNR of 33.35 and an SSIM of 0.92, indicating a favorable protection-fidelity trade-off. Increasing the budget further strengthens protection, reaching the lowest No Key ISM of 0.12 at $\epsilon = 0.08$, while PSNR and SSIM continue to decrease and LPIPS increases. The protection gain largely saturates thereafter, with a No Key ISM of 0.14 at $\epsilon = 0.10$. Fig. 8 shows the same trend, with perturbations becoming progressively more visible as ϵ increases.

For the main experiments, we use $\epsilon = 0.10$ to ensure a fair comparison with prior protection baselines under the same perturbation budget. Nevertheless, the sensitivity results show that strong protection can already be achieved at smaller budgets, with $\epsilon = 0.06$ providing a favorable balance between protection strength and perceptual fidelity.

A.4.2 Sensitivity to Key Length

As defined in Sec. 3.4, the key is represented as $E_k \in \mathbb{R}^{L_k \times d_t}$, where L_k denotes the key length and d_t the text embedding dimension. We study the sensitivity to L_k, which controls the capacity of the key for re-coordinating the disrupted target-side dynamics. We vary $L_k \in \{2, 4, 8, 16, 32\}$ on CelebA-HQ under the standard "a photo of this person" prompt, while keeping all other settings fixed.

Table 7: Sensitivity to key length.

Key Length	No Key				Correct Key			
	FDFR	ISM↓	CLIP-IQA↑	CLIP-Score↑	FDFR	ISM↑	CLIP-IQA↑	CLIP-Score↑
2	0.00	0.14	0.70	26.53	0.00	0.62	**0.77**	24.49
4	0.00	0.14	**0.71**	26.56	0.00	0.70	0.74	24.25
8	0.00	**0.12**	0.70	**26.68**	0.00	0.74	0.71	**24.50**
16 (Ours)	0.00	0.14	**0.71**	26.57	0.00	**0.77**	0.70	24.37
32	0.00	0.14	0.70	26.57	0.00	0.70	0.72	24.43

Overall, key length has little effect on unauthorized protection but directly influences authorized recovery. As shown in Tab. 7, the No Key ISM remains nearly unchanged at 0.12 to 0.14 across all key lengths, indicating that protection is largely insensitive to the capacity of the key. In contrast, Correct Key ISM increases from 0.62 at $L_k = 2$ to 0.70, 0.74, and 0.77 at $L_k = 4, 8, 16$, respectively, showing that increasing key capacity improves recovery up to a moderate length. Increasing L_k further to 32 reduces ISM to 0.70, suggesting that the recovery benefit saturates beyond this point. We therefore use $L_k = 16$ in the main experiments, which provides the strongest authorized recovery while leaving No Key protection essentially unchanged.

A.4.3 Sensitivity to the Attacked Denoising Stage

As described in Sec. 3.5, our optimization focuses on high noise timesteps from the early denoising stage. We examine the sensitivity to this choice by applying the same optimization to three timestep ranges: Early ($t \in \{980, \ldots, 1000\}$), Mid ($t \in \{500, \ldots, 520\}$), and Late ($t \in \{100, \ldots, 120\}$), while keeping all other settings fixed. We evaluate on CelebA-HQ under the standard "a photo of this person" prompt.

Table 8: Sensitivity to the attacked denoising stage.

Phase	No Key				Correct Key			
	FDFR	ISM↓	CLIP-IQA↑	CLIP-Score↑	FDFR	ISM↑	CLIP-IQA↑	CLIP-Score↑
Early (Ours)	0.00	**0.14**	**0.71**	**26.57**	0.00	0.77	**0.70**	24.37
Mid	0.00	0.88	0.57	24.39	0.00	**0.85**	0.69	24.16
Late	0.00	0.91	0.59	24.30	0.00	0.77	0.67	**24.40**

Overall, effective protection depends strongly on targeting the early denoising stage. As shown in Tab. 8, attacking Early timesteps reduces the No Key ISM to 0.14 while retaining a Correct Key ISM of 0.77, producing clear authorization separation. In contrast, attacking Mid or Late timesteps leaves the No Key ISM at 0.88 and 0.91, respectively, close to the Clean level, indicating that these stages provide little leverage for disrupting reference-consistent editing. Mid and Late stage optimization also yields lower CLIP-IQA and CLIP-Score under the No Key condition. These results confirm that the high noise early stage is the most effective region for establishing protected conditional dynamics, supporting our use of $t \in \{980, \ldots, 1000\}$ in the main experiments.

A.4.4 Sensitivity to the Number of Optimization Steps

We study the effect of the optimization budget by varying the number of optimization steps while keeping all other settings fixed. The evaluation is conducted on CelebA-HQ under the standard "a photo of this person" prompt, covering both unauthorized protection and authorized recovery.

Table 9: Sensitivity to the number of optimization steps.

Steps	No Key				Correct Key			
	FDFR	ISM↓	CLIP-IQA↑	CLIP-Score↑	FDFR	ISM↑	CLIP-IQA↑	CLIP-Score↑
Clean	0.00	0.89	0.78	24.13	-	-	-	-
5	0.00	0.81	**0.76**	24.40	0.00	0.91	**0.77**	23.89
20	0.00	0.64	**0.76**	24.88	0.00	0.91	0.75	24.10
50	0.00	0.45	0.73	25.52	0.00	0.82	0.73	24.34
100	0.00	0.26	0.71	26.26	0.00	0.76	0.72	**24.54**
150	0.00	**0.14**	0.71	**26.62**	0.00	0.76	0.70	**24.54**
200 (Ours)	0.00	**0.14**	0.71	26.57	0.00	**0.77**	0.70	24.37

Overall, the effect of optimization steps follows the typical behavior of PGD-based optimization: increasing the number of steps steadily strengthens unauthorized protection before reaching convergence. As shown in Tab. 9, the No Key ISM decreases from 0.81 at 5 steps to 0.26 at 100 steps, and further to 0.14 at 150 steps, with almost no additional improvement at 200 steps. Meanwhile, Correct Key editing remains effective as protection becomes stronger, reaching an ISM of 0.77 at 200 steps. These results indicate that the protection objective largely converges after 150 steps, and we use 200 steps in the main experiments for a stable optimization budget.

A.4.5 Sensitivity to the Number of Sampled Blocks

As described in Sec. 3.5, each optimization iteration samples a subset of transformer blocks $\mathcal{S}$ for computing the steering objectives. We study the sensitivity to the sampling size by varying $|\mathcal{S}| \in \{5, 10, 15, 20, 25\}$ on CelebA-HQ under the standard "a photo of this person" prompt, while keeping all other settings fixed.

Table 10: Sensitivity to the number of sampled blocks.

# Sampled Blocks	No Key				Correct Key			
	FDFR	ISM↓	CLIP-IQA↑	CLIP-Score↑	FDFR	ISM↑	CLIP-IQA↑	CLIP-Score↑
5	0.00	**0.13**	**0.72**	26.63	0.00	0.71	0.71	24.20
10	0.00	0.14	0.71	**26.69**	0.00	0.71	0.70	24.30
15	0.00	0.14	0.71	26.59	0.00	0.75	**0.72**	**24.43**
20	0.00	0.14	0.70	26.54	0.00	0.74	0.71	24.36
25 (Ours)	0.00	0.14	0.71	26.57	0.00	**0.77**	0.70	24.37

Overall, unauthorized protection is highly stable with respect to the number of sampled blocks, while authorized recovery remains effective across different sampling sizes. As shown in Tab. 10, the No Key ISM stays between 0.13 and 0.14 across all settings, indicating that strong protection can be achieved even when only a small subset of blocks is sampled at each iteration. Because blocks are resampled at every iteration, even a small subset can achieve broad coverage over the course of optimization. Correct Key recovery also remains consistently strong, with the highest ISM of 0.77 obtained when sampling 25 blocks. We sample 25 blocks per iteration in the main experiments.

A.5 Robustness to Image Transformations and Purification

We evaluate whether the protected image remains effective after common post-processing and perturbation purification. Specifically, we consider Gaussian corruption, Resize&Crop, and JPEG compression, together with three purification methods: DiffPure (Nie et al., 2022), GrIDPure (Zhao et al., 2024), and IMPRESS (Cao et al., 2023). We evaluate both unauthorized protection under No Key and authorized recovery under the Correct Key on CelebA-HQ using the standard "a photo of this person" prompt.

Table 11: Robustness to common image transformations and purification.

Method	No Key				Correct Key			
	FDFR	ISM↓	CLIP-IQA↑	CLIP-Score↑	FDFR	ISM↑	CLIP-IQA↑	CLIP-Score↑
Clean	0.00	0.89	0.78	24.13	–	–	–	–
Ours	0.00	0.14	0.71	26.57	0.00	0.77	0.70	24.37
Gaussian	0.00	0.17	0.71	26.57	0.00	0.83	0.67	24.24
Resize&Crop	0.00	0.39	0.68	25.07	0.00	0.86	0.67	24.07
JPEG	0.00	0.24	0.71	26.33	0.00	0.81	0.73	24.22
DiffPure	0.00	0.42	0.77	25.60	0.00	0.75	0.71	24.56
GrIDPure	0.00	0.47	0.71	25.00	0.00	0.85	0.65	23.76
IMPRESS	0.00	0.46	0.69	25.20	0.00	0.88	0.61	24.05

Overall, the protected image remains effective under both common transformations and dedicated purification, although stronger purification partially weakens unauthorized protection. As shown in Tab. 11, Gaussian corruption and JPEG compression retain low No Key ISM values of 0.17 and 0.24, respectively, compared with 0.14 without post-processing. Resize&Crop increases No Key ISM to 0.39, while DiffPure, GrIDPure, and IMPRESS further increase it to 0.42, 0.47, and 0.46. Nevertheless, these values remain substantially below the Clean ISM of 0.89, indicating that purification does not fully remove the protection. Meanwhile, Correct Key ISM remains between 0.75 and 0.88 across all transformations and purification methods, showing that authorized recovery is retained after post-processing. These results demonstrate that the proposed protection remains effective under practical image processing and remains resistant to dedicated perturbation removal, while preserving authorized editability.

A.6 Additional Visual Results

We provide additional visual results under diverse held-out prompts, covering interaction, clothing, object manipulation, expression, activity, scene transformation, style transfer, and complex edits.

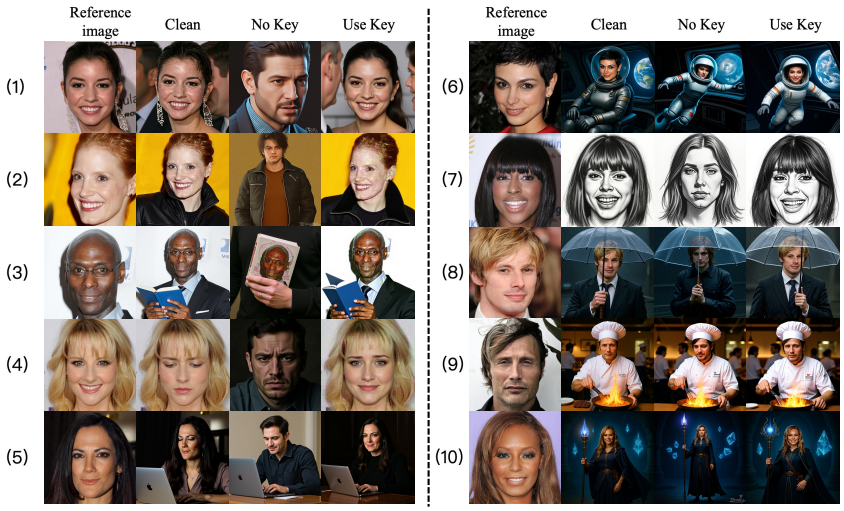


Figure 9: Additional visual results under diverse held-out editing prompts.

The prompts in Fig. 9 are: (1) “Make the person talk to someone,” (2) “Let the person wear a jacket,” (3) “A person holding a book,” (4) “Make the person tired,” (5) “A person using a laptop,” (6) “Transform this person into a space station environment, floating in zero gravity, wearing a futuristic spacesuit, Earth visible through window,” (7) “Draw this person in charcoal sketch style, rough strokes, strong contrast and artistic shading,” (8) “Place this person in heavy rain with wet hair, raindrops on face, holding a transparent umbrella,” (9) “Make this person a professional chef in a busy restaurant kitchen, wearing a white apron and chef’s hat, flames rising from a wok,” and (10) “Dress this person in a fantasy wizard costume with long robe and glowing staff, standing in a magical cave with floating crystals.”

Across these examples, No Key results continue to follow the requested edits while deviating from the protected reference identity, whereas Correct Key results recover reference-consistent editing across both simple and complex prompts.

A.7 Evaluation Prompts

We use 60 held-out prompts for diverse-prompt evaluation, covering attribute-level editing, scene transfer, style transformation, and complex editing. The complete prompt set is listed in Tabs. 12 and 13.

Table 12: Attribute-level editing prompts.

Action	Expression	Gaze
A person sitting down	Make the person smile	A person looking at the camera
A person walking forward	Make the person laugh	A person looking away
A person running	Make the person serious	A person looking down
A person jumping lightly	Make the person tired	A person turning their head
	Make the person confused	
Clothing	**Pose**	**Activity**
Let the person wear a jacket	A person sitting on a chair	A person drinking water
Let the person wear a t-shirt	A person leaning against a wall	A person eating food
Let the person wear a hoodie	A person standing near a window	A person stretching
Let the person wear glasses	A person standing in a doorway	A person relaxing
Let the person wear a backpack		A person thinking
Interaction	**Object**	**Hand Action**
Make the person talk to someone	A person holding a phone	A person raising one hand
Make the person listen to someone	A person holding a book	A person pointing forward
Make the person greet someone	A person holding a cup	A person touching their face
	A person using a laptop	A person holding their hands together
	A person taking a photo	A person gesturing while speaking

Table 13: Scene, style, and complex editing prompts.

Scene Transfer	Style Fusion	Complex Edit
Place this person on a tropical beach with palm trees, ocean waves, and golden sunset	Render this person in an oil painting style with visible brushstrokes and rich textured colors	Transform this person into a space station environment, floating in zero gravity, wearing a futuristic spacesuit, Earth visible through window
Put this person in a snowy mountain landscape with pine trees and fresh powder snow	Capture this person in black and white film noir style, dramatic shadows and high contrast	Dress this person as a medieval knight with sword and shield, standing in a castle courtyard with banners and dramatic sky
Move this person to a bustling city street with skyscrapers, taxis, and crowds	Place this person in heavy rain with wet hair, raindrops on face, holding a transparent umbrella	Make this person a professional chef in a busy restaurant kitchen, wearing a white apron and chef's hat, flames rising from a wok
Put this person inside a cozy medieval tavern with wooden tables, candlelight, and stone walls	Put this person in thick fog or mist, blurred silhouette, soft diffused light and low visibility	Place this person in a hospital setting, wearing a white doctor's coat and stethoscope, standing in a bright operating room with surgical lights
Place this person in a futuristic cyberpunk city with flying cars and holographic billboards	Render this person in a cyberpunk holographic projection style with glitch effects and neon overlays	Dress this person in a fantasy wizard costume with long robe and glowing staff, standing in a magical cave with floating crystals
Put this person in a quiet library surrounded by tall bookshelves and warm reading lamps	Draw this person in charcoal sketch style, rough strokes, strong contrast and artistic shading	Put this person on a submarine deep-sea viewing deck, giant squid visible through porthole, wearing a naval officer's uniform
Place this person in a dense enchanted forest with glowing mushrooms and twisted ancient trees		
Put this person in a blooming cherry blossom tunnel in spring, pink petals falling around		