DECONTEXT AS DEFENSE: SAFE IMAGE EDITING IN DIFFUSION TRANSFORMERS

Anonymous authors
Paper under double-blind review

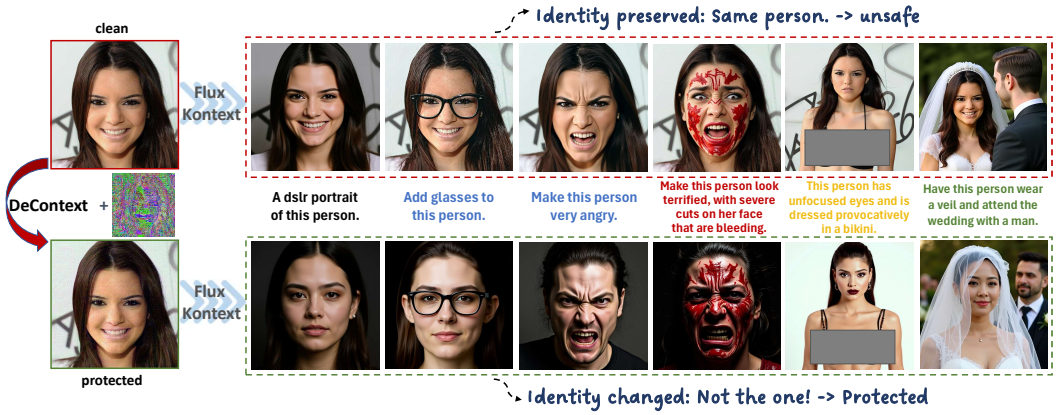


Figure 1: Overview of our protection method against malicious edits by Flux-Kontext. A clean user image (top left) can be altered for neutral, violent, sexual, or misleading edits. Our **DeContext** injects imperceptible perturbations into the input image, preventing identity preservation in edited results while retaining visual quality.

ABSTRACT

In-context diffusion models allow users to modify images with remarkable ease and realism. However, the same power raises serious privacy concerns: personal images can be easily manipulated for identity impersonation, misinformation, or other malicious uses, all without the owner's consent. While prior work has explored safeguarding against misuse in personalized text-to-image generation, the robustness of modern, large-scale in-context DiT-based models remains largely unexamined. In this paper, we propose **DeContext**, a new method to safeguard input images from unauthorized in-context editing. Our key insight is that contextual information from the source image propagates to the output primarily through multimodal attention layers. By injecting small, targeted perturbations that weaken these cross-attention pathways, DeContext breaks this flow, effectively decoupling the link between input and output. This simple defense is both efficient and robust. We further show that early denoising steps and specific transformer blocks dominate context propagation, which allows us to concentrate perturbations where they matter most. Experiments on *Flux Kontext* and *Step1X-Edit* show that DeContext consistently blocks unwanted image edits while preserving visual quality. These results highlight the effectiveness of attention-based perturbations as a powerful defense against image manipulation.

1 INTRODUCTION

Large-scale diffusion models now dominate visual generation. Built on diffusion transformer (DiT) architectures and trained at scale, they produce high-fidelity, diverse images (Podell et al., 2024; Labs, 2024; Esser et al., 2024). This progress has created a growing demand for personalization and controllable generation, which allows users to adapt models to novel concepts (Gal et al., 2022; Ruiz et al., 2023) or editing instructions from a few examples (Zhang et al., 2023; Ye et al., 2023; Bar-Tal et al., 2023; Qin et al., 2023). Very recently, a new paradigm emerges: editing as in-context learning. Models such as Flux Kontext (Labs et al., 2025) and Qwen-Image (Wu et al., 2025) integrate user

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.

control directly into DiT-style networks via attention mechanisms, enabling a more scalable and flexible approach to controllable synthesis.

However, the very strength of in-context personalization also introduces serious privacy risks. Since they operate at inference time and require only a single image, adversaries can easily exploit publicly shared photos to clone identities, mimic artistic styles, or reproduce proprietary content without consent. Consequently, as shown in Fig. 1, the images users share online to express themselves may be weaponized to generate deepfakes, impersonation, and copyright violation, entirely bypassing conventional data protection mechanism.

To reduce these risks, prior studies have proposed defense strategies against unauthorized image editing. Most methods target either general text-to-image models (Liu et al., 2024a; 2025a; Zhang et al., 2025) or traditional optimization-based personalization (Van Le et al., 2023; Liang & Wu, 2023; Xu et al., 2024a;b; Liu et al., 2024b; Wang et al., 2024a). The most related approaches, such as PhotoGuard (Salman et al., 2023), Editshield (Chen et al., 2024a) and FaceLock (Wang et al., 2025), are designed for image-to-image personalization.

However, these current protections are insufficient for the emerging threats we confront. They were primarily designed for legacy UNet-based architectures and are useless when employed with modern Transformer-based in-context learning models. These solutions significantly compromise the fidelity-protection trade-off. Even when altered, they generate considerable visual distortions while attempting to impede the objective concept. This underscores a notable shortcoming: there is currently *no sufficiently effective safeguard* specifically tailored for these innovative DiT-based image editing models.

This paper proposes **DeContext**, a specialized defense mechanism to protect input images from illicit DiT-based in-context modifications. Our methodology is founded on a crucial understanding that, in DiT models, the impact of a conditional image disseminates exclusively via *multi-modal attention* (introduced in Sec.2.1). We discovered that by interrupting this attention flow, we can inhibit unlawful identity or style transfer without compromising overall image quality.

Building on this finding, DeContext introduces precisely localized, attention-aware perturbations to the input image. These are strategically designed to suppress attention activations between target queries (the generated image) and context keys (the private image) to minimize the interaction. Guided by our analysis, we selectively apply these perturbations at early denoising steps and within the most influential, early-to-middle transformer blocks. This design effectively **De**taches the editing **Context** from the generated image, preserving visual fidelity without any model modification.

Through extensive experiments, we demonstrate that DeContext provides robust protection against malicious editing in DiT-based in-context personalization, outperforming prior defenses on both protection effectiveness and image quality. Images protected by DeContext consistently foil in-context edits and identity extraction, leading to over 70% drop in face recognition accuracy, while preserving image quality typically within 20% of the clean images.

Our contributions are summarized as follows: (1) We introduce the first defense framework for preventing unauthorized in-context image editing in DiT-based models. (2) We identify the multi-modal attention mechanism as the key vulnerability that enables DiT-based editors to leak and propagate contextual information from the conditioning input into the generated image. (3) We propose **DeContext**, a targeted perturbation strategy that disrupts this attention flow. By further concentrating perturbations on the identified early denoising timesteps and influential transformer blocks, we largely enhance the protection efficacy in practice. (4) Our experiments on FLUX Kontext and Step1X-Edit show that DeContext effectively removes contextual information while preserving high visual fidelity in the outputs generated from protected contexts.

2 Motivation Analysis

Our goal is to investigate the robustness of conditional Diffusion Transformers (DiTs) by attacking their ability to use context images. Our key insight comes from two experiments: standard adversarial attacks fail, but directly disrupting the model's internal attention mechanism successfully removes the context's influence. This motivates our proposed method: a new defense specifically designed to disrupt this attention-based conditioning mechanism.

2.1 BACKGROUND

Diffusion Transformer with Conditions. Recent diffusion transformers (DiTs) replace the UNet backbone with a transformer that jointly processes text and image tokens. Following FLUX-Kontext (Labs et al., 2025), we consider image generation conditioned on a text prompt c and context image y, where both x and y are encoded by a frozen VAE $\mathcal{E}_{\text{vae}}$ and c by a pretrained language encoder. The model approximates the conditional distribution $p_\theta(x \mid y, c)$ via a rectified flow-matching objective:

$$\mathcal{L}_\theta = \mathbb{E}_{t,x,y,c}\|v_\theta(z_t, t, y, c) - (\epsilon - x)\|_2^2, \tag{1}$$

where $z_t = (1-t)x + t\epsilon$, and $\epsilon \sim \mathcal{N}(0, I)$. DiTs integrate multi-modal information through multi-modal attention (MMA) by concatenating tokens from text, target, and context:

$$\mathbf{Z} = [\,\mathbf{Z}_{\text{text}}\,;\,\mathbf{Z}_{\text{tgt}}\,;\,\mathbf{Z}_{\text{ctx}}\,]. \tag{2}$$

This concatenation and subsequent attention is the core mechanism by which the model learns to reflect the information from the context y via $\mathbf{Z}_{\text{ctx}}$ onto target x via $\mathbf{Z}_{\text{tgt}}$. As we will show, this mechanism is central to both our analysis and our proposed attack in Section 3.

Adversarial Attack. Adversarial attacks aim to find imperceptible noise that alter a model's prediction. Typical formulations focus on classification or image generation attack: for a model f and an input image x, an adversarial example x' is crafted to remain visually similar to x while causing a misprediction $y_{\text{true}} \neq f(x')$ (untargeted), or to force a predefined prediction $y_{\text{target}} = f(x') \neq y_{\text{true}}$ (targeted). The perturbation is usually constrained within an η-ball under an ℓ_p metric, i.e. $\|x' - x\|_p \leq \eta$. Denoting $\Delta = \{\delta : \|\delta\|_p \leq \eta\}$, the untargeted and targeted optimization can be written as

$$\delta_{\text{adv}} = \arg\max_{\delta\in\Delta} \mathcal{L}\big(f(x+\delta),\, y_{\text{true}}\big), \tag{3}$$

$$\delta_{\text{adv}} = \arg\min_{\delta\in\Delta} \mathcal{L}\big(f(x+\delta),\, y_{\text{target}}\big). \tag{4}$$

These objectives are solved iteratively. The standard tool is Projected Gradient Descent (PGD) (Madry et al., 2017). It takes gradient steps followed by projection onto the norm ball. The update for untargeted attack is:

$$x'_0 = x, \tag{5}$$

$$x'_k = \Pi_{x,\eta}\big(x'_{k-1} + \alpha \cdot \text{sign}(\nabla_x \mathcal{L}(f(x'_{k-1}), y)\big), \tag{6}$$

where $\Pi_{x,\eta}$ projects onto the ℓ_p ball around x. We review the related work in Appendix B.

2.2 CONTEXT PROPAGATES THROUGH ATTENTION

Based on the PGD attacks and MMA-based conditioning, we next investigate how a direct attack on the diffusion transformer performs. Our findings are summarized below.



Figure 2: Attack results. Standard attack only yields re-lighting artifacts. Removing context(Q)→target(K) attention has minimal effect, while removing target(Q)→context(K) detaches context.

Study I: Failure of the Standard Attack. We first test the most direct choice: applying PGD to the standard diffusion training objective. Specifically, we perform an untargeted attack (Eq. 3) that maximizes the flow-matching loss (Eq. 1) of Flux-Kontext. We refer to this baseline as **Diff-PGD**.

Unfortunately, this approach is ineffective. As shown in Fig. 2 (second column), the attack introduces only mild blur and re-lighting artifacts, while the facial identity and overall structure remain clearly preserved. This result shows that simply attacking the end-to-end flow-matching loss is insufficient to suppress contextual information. Despite the perturbation, the model can still recover substantial information from the context image.

Study II: Attention Intervention Analysis. The failure of Diff-PGD motivates us to examine how contextual information is propagated inside the model. In particular, we focus on multi-modal

attention (Eq. 2), which directly mediates interactions between the context and target representations. We therefore perform controlled interventions during inference by selectively removing specific cross-image attention connections.

We consider two settings. (1) *Context image as query, target image as key*. As shown in Fig. 2 (second-right column), discarding this attention region does not substantially alter the generated outputs. This observation indicates that suppressing this interaction fails to block the private image's influence, leading to an ineffective defense. (2) *Target image as query, context image as key*. The outcome, depicted in Fig. 2 (rightmost), demonstrates that the intervention successfully severs context propagation, resulting in generation solely influenced by the text prompt. These findings suggest that the critical pathway for context leakage lies in the target-to-context attention flow.

3 DECONTEXT

Based on the analysis above, this section presents *DeContext*, a new method to protect the privacy of user-supplied condition images in DiT-based in-context editing. We first describe the overall context-detach mechanism. Then, guided by the context propagation analysis, we perform attack on the most influential timesteps and transformer blocks to maximize protection. The detailed algorithm will be provided in Appendix F.2.

3.1 CONTEXT DETACHMENT VIA ATTENTION

Our goal is to prevent the target image tokens from "*attending*" to the context image tokens. We achieve this by directly minimizing the cross-attention weights between them. Surprisingly, as shown in Sec. 2.2, this simple objective alone is sufficient to decouple the target from the context and prevent private information from propagating into the generation.

Context Attention Suppression Objective. Within a transformer block, we consider the query q and key k embeddings, $q, k \in \mathbb{R}^{H\times S\times d}$, where H is head counts, S is total token sequence length covering text, target, and context tokens, and d is the hidden dimension. For efficiency, We consider attention computation to only the rows of the attention map corresponding to target-image queries Q:

$$A_{q,:} = \mathrm{Softmax}\big((qk^\top)/\sqrt{d}\big), \quad q \in Q. \tag{7}$$

To measure how much the target attends to the context, we define the *context proportion* r_{ctx} as the average attention from target queries to context keys:

$$r_{\text{ctx}} = \frac{1}{HB|Q|} \sum_{h,b} \sum_{q\in Q} \sum_{k\in\mathcal{C}} A_{qk}^{(h,b)}, \tag{8}$$

where B denotes the number of blocks. This ratio r_{ctx} is a direct measure of how strongly the model uses the context image to generate the target. To "detach" the context, our adversarial objective is to maximize the following loss function, which forces r_{ctx} towards zero:

$$\mathcal{L}_{\text{DeContext}} = 1 - r_{\text{ctx}}, \tag{9}$$

We use gradient ascent to iteratively update the pixels of the context image x to maximize our loss:

$$x^{\text{adv}} \leftarrow \mathrm{clip}\big(x^{\text{adv}} + \alpha\ \mathrm{sign}(\nabla_{x^{\text{adv}}}\mathcal{L}),\ x-\epsilon,\ x+\epsilon\big). \tag{10}$$

As Figure 3 illustrates, all model weights are frozen; only the context image is updated through projected gradient ascent.

Robust Attack via Sampling. A successful attack should remain robust across different generation conditions. The generated output depends on the text prompt p, diffusion timestep t, and noise z. Optimizing the perturbation for a single combination may therefore overfit to that specific condition, while jointly optimizing over all possible combinations is computationally prohibitive.

To implement this, we maintain a pool $\mathcal{P}$ of 60 editing commands and a target timestep interval (selection criteria will be introduced in Section 3.2). During each attack iteration, DeContext randomly samples one prompt $p_i \sim \mathcal{P}$, a diffusion timestep $t \sim \mathcal{T}$, and a noise z, then computes the loss $\mathcal{L}(c; p_i, t, z)$ and updates the perturbed context image via gradient ascent. This single-sample gradient provides an unbiased approximation of the full expected objective:

$$\nabla_c\, \mathbb{E}_{p,t,z}\, \mathcal{L}(c; p, t, z) \approx \mathbb{E}_{p,t,z}\, \nabla_c\, \mathcal{L}(c; p, t, z). \tag{11}$$

so the learned perturbation is effective across prompts and seeds.

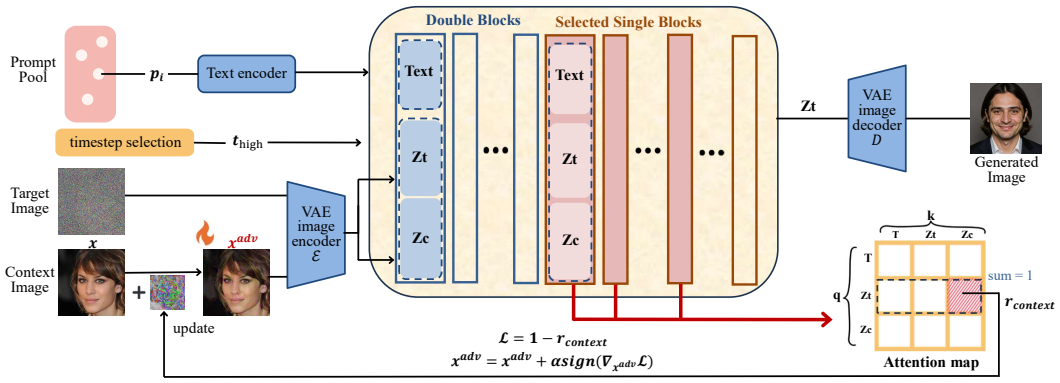
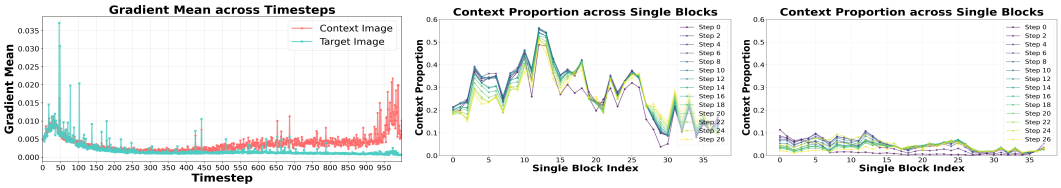


Figure 3: Overview of our DeContext pipeline. Given a prompt, timestep, noisy target, and context image, DeContext perturbs the context to suppress its attention in the diffusion model. Iterative gradient updates minimize attention activation, detaching the context from influencing generation.

3.2 Concentrated Context Detachment

Our goal is to detach the target generation from the context image. However, we find that context influence is not uniform across the diffusion transformer. Instead, it is concentrated at particular timesteps and transformer blocks. This motivates *concentrated detachment*, where we optimize only at the locations where the context has the strongest influence.



(a) Time-wise gradients

(b) Block-wise attention (Before attack)

(c) Block-wise attention (After attack)

Figure 4: Timesteps *(a)* and blocks *(b)(c)* analysis. Gradients of context image dominate at high timesteps (early denoising). Context proportion is high in early-to-mid blocks in DiT.

Timestep-wise Detachment.

Analysis. We first ask *when* the context image has the strongest influence during denoising. For a controlled comparison, we set the target image identical to the context image and perform standard denoising using the prompt "a photo of this person".

To measure this, at each timestep t, we compute the mean absolute gradients with respect to the target $\mathbf{Z} * \text{tgt}$ and the context image $\mathbf{Z} * \text{ctx}$. These gradients quantify how strongly the prediction depends on the target and context, respectively.

$$g_t^{\mathbf{X}} = \frac{1}{|\mathbf{X}|} \sum_i \left| \frac{\partial \mathcal{L}}{\partial \mathbf{X}_i} \right|, \quad \mathbf{X} \in \{\mathbf{Z}_{\text{tgt}}, \mathbf{Z}_{\text{ctx}}\}, \tag{12}$$

We repeat this process across all timesteps and average over several samples for stability. As shown in Fig. 4a, early timesteps (large t, high noise) exhibit higher gradients for $\mathbf{Z}_{\text{ctx}}$, while later timesteps (small t, low noise) show dominant gradients for $\mathbf{Z}_{\text{tgt}}$, indicating that the context primarily guides the denoising process in its early phase of generation.

Strategy. Based on this finding, we restrict our optimization to these early, influential timesteps $t_i \in [t_{\text{high}}, T]$. At these high-noise steps, we also approximate the target image with random Gaussian noise $z \sim \mathcal{N}(0, I)$, which avoids the need for paired images and simplifies the optimization.

Block-wise Detachment.

Analysis. Next, we ask *where* context information is most strongly accessed within the transformer. Following Eq. 8, we compute the context proportion r_{ctx} for each transformer block, which measures how much attention from target queries is assigned to context keys.

As shown in Fig. 4b, r_{ctx} is substantially higher in the early-to-middle blocks. These blocks therefore play a larger role in retrieving information from the context image.

Strategy. Therefore, we focus our suppression strategies on these specific, context-heavy transformer blocks[1]. As shown in Fig. 4c, this strategy significantly suppresses r_{ctx} across all blocks. Notably, although later blocks are not directly targeted, perturbations applied to earlier blocks propagate through the network and indirectly suppress their attention. This confirms that our concentrated attack effectively disrupts the propagation of context information, achieving the intended detachment.

4 Experiments

4.1 Experimental Setup

Defense Objective. Prior defenses (Van Le et al., 2023; Liang & Wu, 2023; Xu et al., 2024a;b; Liu et al., 2024b) often protect privacy by degrading the visual fidelity of generated output. In contrast, DeContext removes identity information from the context while keeping the generation realistic and consistent with the editing prompt. This preserves the model's normal editing behavior while protecting the private identity.

Datasets. Experiments are conducted on high-quality face images derived from two datasets: VGGFace2 (Cao et al., 2018) and CelebA-HQ (Liu et al., 2015).

For each dataset, we work with a subset containing 50 distinct identities. The original images are cropped and resized 512×512 for convenient verification. For T2I baselines, we select four images per identity for DreamBooth fine-tuning, following their default settings. For I2I tasks, the same single image per identity (one of the four used for T2I personalization) is used as context input to the diffusion model for consistent evaluation.

Target Models. Our primary evaluation target is FLUX.1-Kontext-dev (Labs et al., 2025), a state-of-the-art DiT-based in-context editing model. We also evaluate generalization on Step1X-Edit (Liu et al., 2025b) (See Sec. 4.4).

Baselines. Since no prior work directly addresses DiT-based I2I editing defense, we adapt existing protection methods:

- *T2I personalization defenses:* Anti-DB (Van Le et al., 2023), AdvDM (Liang et al., 2023), and CAAT (Xu et al., 2024a) applied to DreamBooth followed their default settings.
- *I2I defense:* FaceLock (Wang et al., 2025) for InstructPix2Pix (Brooks et al., 2023).
- *Naive baseline:* Diff-PGD maximizing reconstruction loss. The working principle is discussed in Sec. 2.2 and detailed in Appendix F.1.

Implementation Details. We set the step size to $\alpha = 0.005$ and the noise budget to $\eta = 0.1$, running a total of 800 optimization steps. At each training iteration, we randomly sample a diffusion timestep t from the interval $\{980, \ldots, 1000\}$, and a prompt from a fixed pool of 60 prompts (complete list provided in the Appendix F.3). Experiments are conducted on a single NVIDIA A800 80 GB GPU. Unless otherwise specified, these settings remain consistent across all experiments.

Comparison with protection baselines. To evaluate whether DeContext provides stronger protection for user image privacy, we compare it against prior UNet-based defenses under the same setup. Clean experiments are conducted on Flux Kontext and we rerun all baselines under noise budget $\eta = 0.1$ for a fair comparison. For T2I personalization on DreamBooth, we follow Anti-DreamBooth (Van Le et al., 2023) prompts "a photo of sks person" and "a dslr portrait of sks person", where *sks* is the keyword. For I2I tasks, prompts are adapted to "a photo of this person" and "a dslr portrait of this person". These prompts are excluded in our 60-prompt training pool to show that DeContext generalizes to test prompts. Each identity–prompt pair generates 30 samples.

[1] Implementation details are discussed in Sec. 4.3

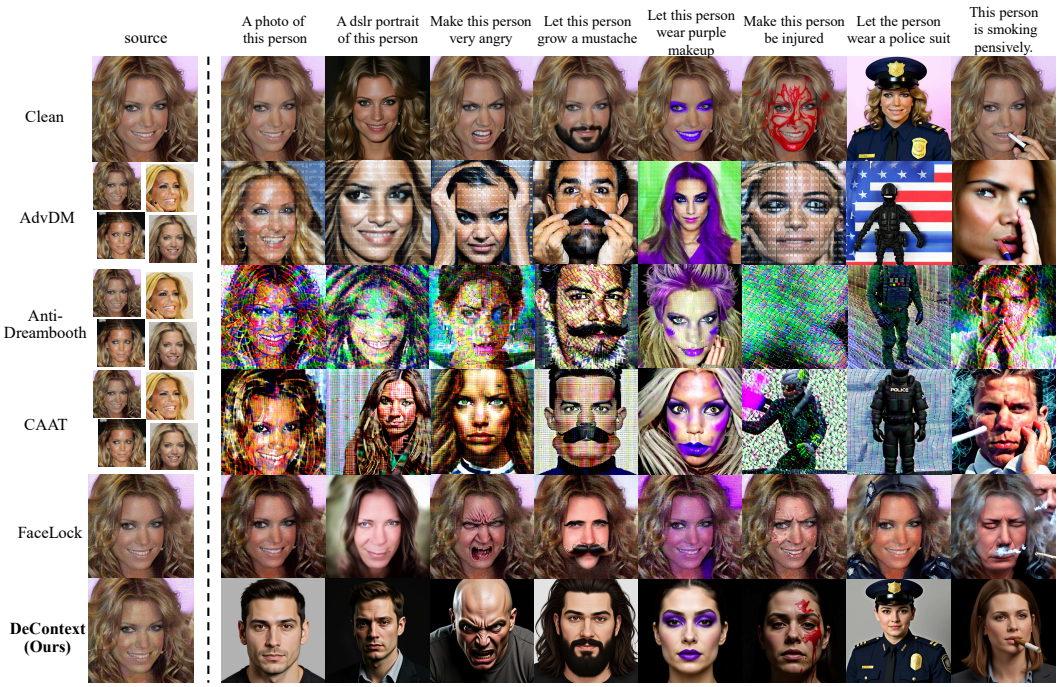


Figure 5: Defense Comparison. Previous attacks on T2I fine-tuned models and I2I models introduce visual artifacts in the generated images, whereas DeContext effectively removes context identity while preserving high visual quality.

As shown in Tab. 1, DeContext generally outperforms prior UNet-based defenses. In terms of identity removal, DeContext detects more faces while more effectively eliminating identifiable information. On CelebA-HQ using the prompt "a photo of this person", it achieves an ISM of 0.12, markedly better than the best baseline (0.32). CLIP-I scores also remain consistently lower than those of the state-of-the-art I2I defense FaceLock (Wang et al., 2025). For image quality, DeContext also achieves lower BRISQUE and FID, producing fewer artifacts and more natural outputs. Note that, the Diff-PGD baseline performs poorly at context detachment, with notably high ISM and CLIP-I scores.

As shown in Fig. 5, prior T2I defenses (Anti-DreamBooth, AdvDM, CAAT) introduce severe distortions with colorful noise and corrupted textures, while FaceLock produces color shifts and sometimes fails to fully remove identity. DeContext, in contrast, consistently generates realistic outputs that are completely unrelated to the source identity across all editing instructions, achieving the best identity detachment with a favorable protection–quality trade-off.

Evaluation metrics. Following prior text-to-image personalization and editing works, we use the same evaluation protocol for fair baseline comparison. We evaluate generated images under protected contexts from two aspects: *whether the identity is altered* and *how well the image quality is preserved*.

- **Identity Protection.** We use ***Face Detection Failure Rate (FDFR)*** to measure the percentage of samples where RetinaFace (Deng et al., 2020) fails to detect a face. Once detected, we use ***Identity Score Matching (ISM)*** to compute the distance between ArcFace (Deng et al., 2019) embeddings (smaller is better). For I2I tasks, we additionally compute ***CLIP Image Similarity (CLIP-I)***, which measures semantic similarity between source and generated images in feature space.

- **Image Quality.** measured by ***SER-FQA*** (Terhorst et al., 2020) (face-specific), ***BRISQUE*** (Mittal et al., 2012) (perceptual) and ***FID*** (statistical realism). Since we aim to remove identity cues while keeping realism, higher SER-FQA, lower BRISQUE and lower FID denote better results.

While we mainly evaluate on faces as the most representative and safety-critical example, we additionally include object-level results in the Appendix D.

Dataset	Method	I2I	"a photo of *this* person"					"a dslr portrait of *this* person"				
			FDFR↓	ISM↓	CLIP-I↓	BRISQUE↓	FID↓	FDFR↓	ISM↓	CLIP-I↓	BRISQUE↓	FID↓
VGGFace2	Clean	✓	0.00	0.78	0.96	15.78	200.34	0.00	0.71	0.88	14.22	183.33
	Anti-DB (Van Le et al., 2023)	✗	0.76	0.21	-	37.33	421.70	0.86	0.23	-	40.92	483.54
	Adv-DM (Liang et al., 2023)	✗	0.30	0.20	-	35.84	465.58	0.11	**0.10**	-	**30.37**	280.18
	CAAT (Xu et al., 2024a)	✗	0.80	0.17	-	39.44	428.80	0.86	0.17	-	39.11	429.96
	FaceLock (Wang et al., 2025)	✓	0.01	0.19	0.54	30.75	233.42	0.01	0.16	0.59	40.65	303.10
	Diff-PGD	✓	0.00	0.60	0.76	74.19	253.85	0.00	0.58	0.77	45.91	213.57
	DeContext	✓	**0.00**	**0.16**	**0.50**	**23.80**	**210.98**	**0.00**	0.23	**0.56**	36.83	**207.19**
CelebA-HQ	Clean	✓	0.00	0.79	0.95	14.20	130.01	0.00	0.61	0.82	12.00	139.22
	Anti-DB (Van Le et al., 2023)	✗	0.73	0.36	-	38.83	395.27	0.74	0.27	-	38.99	424.60
	AdvDM (Liang et al., 2023)	✗	0.07	0.35	-	22.67	280.91	0.03	0.22	-	**28.79**	231.49
	CAAT (Xu et al., 2024a)	✗	0.60	0.32	-	46.01	413.68	0.78	0.26	-	40.16	370.56
	FaceLock (Wang et al., 2025)	✓	0.00	0.51	0.73	29.95	**200.87**	0.00	0.40	0.65	40.96	231.15
	Diff-PGD	✓	0.01	0.68	0.80	74.30	235.89	0.00	0.59	0.79	38.96	230.72
	DeContext	✓	**0.00**	**0.12**	**0.51**	**20.56**	229.68	**0.00**	**0.20**	**0.58**	39.17	**209.37**

Table 1: Quantitative results on VGGFace2 and CelebA-HQ datasets.[2]

Method	"Make this person very angry."					"This person is wearing a smoky eye makeup."				
	ISM↓	CLIP-I↓	FID↓	BRISQUE↓	SER-FQA↑	ISM↓	CLIP-I↓	FID↓	BRISQUE↓	SER-FQA↑
Clean	0.60	0.88	139.41	13.67	0.78	0.75	0.91	141.28	13.77	0.74
DeContext	0.07	0.49	252.44	33.17	0.79	0.21	0.58	202.77	14.10	0.60
Method	**"This person is looking at the mirror."**					**"Add glasses to this person."**				
	ISM↓	CLIP-I↓	FID↓	BRISQUE↓	SER-FQA↑	ISM↓	CLIP-I↓	FID↓	BRISQUE↓	SER-FQA↑
Clean	0.72	0.85	190.73	13.12	0.72	0.60	0.88	179.71	15.00	0.73
DeContext	0.36	0.61	241.49	10.55	0.67	0.14	0.57	212.45	16.24	0.76

Table 2: Quantitative results under different prompts.

4.2 Main Results

Generalization to multiple face editing prompts. We further evaluate the robustness of DeContext under diverse editing prompts. Specifically, we test it on CelebA-HQ using four additional in-context editing instructions. For each prompt, we sample 30 distinct identities and generate 10 images per identity. With quantitative results in Tab. 2, identity metrics show notable reductions across all prompts: ISM and CLIP-I decrease by 73% and 36% on average. Image quality remains stable, with BRISQUE and SER-FQA varying under 10% for most cases. Since the real FID embeddings are derived from clean faces, they align closely with clean outputs but not with perturbed ones, leading to relatively higher FID. Qualitative results in Fig. 5 visually show that other baselines tend to preserve identity to some extent and often produce visible artifacts. In contrast, DeContext achieves a clear trade-off between identity removal and overall visual quality.

4.3 Ablation studies

To evaluate the impact of the attack budget and our three key designs (attention mechanism, block-wise and time-wise detachment), we conduct four ablations following the setup in Sec. 4.1.

Ablation on attack budget. We examine how performance varies with the attack's norm ball size. As shown in Tab. 3, a larger budget η improves identity detachment but introduces more visual artifacts and slightly reduces image quality. We use $\eta = 0.1$ as the default.

Ablation on attention mechanism. We validate the necessity of attacking cross-attention via the Diff-PGD baseline. As shown in Fig. 2 and Tab. 1, maximizing diffusion training loss only introduces minor artifacts and fails to block editing, indicating that attacking cross-attention is essential for modern DiTs.

Ablation on diffusion timesteps. We evaluate four timestep settings: 980–1000 (default), 0–1000 (all), 0–500, and 500–1000. As shown in Tab. 4, performance improves as perturbations are applied to larger timesteps, with the best results achieved at 980–1000. This is consistent with our timestep-wise analysis (Sec. 3.2) that early denoising steps guide context propagation, making perturbations at these timesteps most effective.

[2]Baselines are applied to their default models with their default settings (with $\eta = 0.1$);

Dataset	η	"a photo of *this* person"				
		ISM↓	CLIP-I↓	SER-FQA↑	BRISQUE↓	FID↓
VGGFace2	0.00	0.78	0.96	0.76	15.78	200.34
	0.05	0.40	0.67	**0.71**	**20.05**	**197.86**
	0.10*	0.16	0.51	0.65	23.80	210.98
	0.15	**0.14**	**0.50**	0.65	24.17	213.54
CelebA-HQ	0.00	0.79	0.95	0.76	14.20	130.01
	0.05	0.36	0.66	**0.69**	**15.38**	**177.95**
	0.10*	0.12	0.51	0.64	20.56	229.68
	0.15	**0.11**	**0.50**	0.64	22.87	228.72

Table 3: Ablation on attack budgets η. (* denotes default)

Metrics	Diffusion Timesteps				
	Clean	980–1000*	0–1000	0–500	500–1000
ISM↓	0.79	**0.13**	0.56	0.67	0.45
CLIP-I↓	0.95	**0.49**	0.80	0.87	0.74

Table 4: Ablation on timesteps.

Metrics	Perturbed Blocks		
	Double All	Single 0–25*	Single 12–37
ISM ↓	0.52	0.13	0.20
CLIP-I ↓	0.73	0.49	0.57

Table 5: Ablation on blocks.

Dataset	Method	"a photo of this person"					"a dslr portrait of this person"				
		ISM↓	CLIP-I↓	FID↓	BRISQUE↓	SER-FQA↑	ISM↓	CLIP-I↓	FID↓	BRISQUE↓	SER-FQA↑
CelebA-HQ	Clean	0.62	0.92	125.64	10.82	0.76	0.67	0.90	166.49	8.51	0.73
	DeContext	0.13	0.56	160.87	18.89	0.81	0.16	0.60	183.37	13.84	0.77
	Method	"This person is wearing a smoky eye makeup."					"Give this person a beard."				
		ISM↓	CLIP-I↓	FID↓	BRISQUE↓	SER-FQA↑	ISM↓	CLIP-I↓	FID↓	BRISQUE↓	SER-FQA↑
	Clean	0.67	0.92	134.42	8.87	0.74	0.57	0.76	161.25	5.28	0.67
	DeContext	0.17	0.59	161.89	15.45	0.72	0.08	0.50	236.15	15.69	0.70

Table 6: Evaluations on Step1X-Edit.

Ablation on perturbed blocks. Flux-Kontext consists of 19 double blocks followed by 38 single blocks. To validate our finding and layer selection in Sec. 3.2, we test three settings: attacking all double blocks, the first 25 single blocks, and the last 25 single blocks (Tab. 5). Consistent with our block-wise analysis, early-to-mid single blocks play a key role in context propagation. Perturbing only these blocks achieves the best trade-off between identity detachment and efficiency.

4.4 Extension Results on Step1X-Edit

Step1X-Edit (Liu et al., 2025b) is another DiT-based model designed for image editing. To assess the generalization capability of our proposed attack, we directly apply DeContext to Step1X-Edit using the same experimental setup as described for FLUX.1-Kontext in Section 4.1. Specifically, we evaluate on the CelebA-HQ (Liu et al., 2015) dataset, generating 50 images per identity-prompt pair.



Figure 6: Qualitative results on Step1X-Edit.

As shown in Fig. 6, DeContext reliably avoids identity transfer while producing realistic results across a wide range of editing directions from neutral prompts "dslr portrait") to specific appearance changes ("smoky eye", "add beard"). Quantitatively, as summarized in Tab. 6, DeContext achieves comparable identity removal and image quality performance to that observed on FLUX.1-Kontext, with a notable average ISM reduction of more than 80%. These results demonstrate that DeContext maintains its effectiveness across different DiT-based image-to-image architectures.

5 Conclusion and Future Work

In this paper, we present DeContext, the first defense for DiT-based in-context image editing. DeContext protects user images by disrupting the propagation of context through multi-modal attention. Experiments on FLUX.1-Kontext and Step1X-Edit demonstrate that DeContext effectively detaches the contextual information from the input while preserving high quality. In future work, we aim to improve the adaptability of our defense to continuously evolving large-scale generative models and enhance its robustness for real-world protection.

AI USE STATEMENT

Generative AI tools supported manuscript review, experimental planning, code development, and language editing. The authors conducted the experiments, verified the code, results, text, and references, and take full responsibility for the final manuscript.

ETHICS STATEMENT

This work explores image protection against unwanted identity-preserving editing, aiming to support privacy and user control over personal images. All participants in the user study described in Appendix G participated voluntarily.

REPRODUCIBILITY STATEMENT

Section 3 describes the proposed method and optimization objective. Section 4.1 specifies datasets, baselines, perturbation budgets, and evaluation metrics. Appendix F provides the algorithms and optimization prompt pool. Sections 4.3 and 4.4 report ablations and Step1X-Edit experiments. Appendices C, D, and G detail preprocessing evaluations, additional image domains, and the user study.

REFERENCES

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22669–22679. IEEE, 2023.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402. IEEE, 2023.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22503–22513. IEEE, 2023.

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. IEEE, 2018.

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

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, 2024a.

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6593–6602. IEEE, 2024b.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

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

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in neural information processing systems*, 34:19822–19835, 2021.

Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M Roy. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using b-lora. In *European Conference on Computer Vision*, pp. 181–198. Springer, 2024.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017.

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.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. Unlearnable examples: Making personal data unexploitable. *arXiv preprint arXiv:2101.04898*, 2021.

Weiyao Huang, Liqin Wang, Ziqi Sheng, and Wei Lu. Nulledit: Stealthy image protection via vlm condition redirection. *arXiv preprint arXiv:2608.10870*, 2026.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24174–24184. IEEE, 2024.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1931–1941. IEEE, 2023.

Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

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.

Chumeng Liang and Xiaoyu Wu. Mist: Towards improved adversarial examples for diffusion models. *arXiv preprint arXiv:2305.12683*, 2023.

Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. *arXiv preprint arXiv:2302.04578*, 2023.

Gongye Liu, Menghan Xia, Yong Zhang, Haoxin Chen, Jinbo Xing, Yibo Wang, Xintao Wang, Yujiu Yang, and Ying Shan. Stylecrafter: Enhancing stylized text-to-video generation with style adapter. *arXiv preprint arXiv:2312.00330*, 2023.

Runtao Liu, Ashkan Khakzar, Jindong Gu, Qifeng Chen, Philip Torr, and Fabio Pizzati. Latent guard: a safety framework for text-to-image generation. In *European Conference on Computer Vision*, pp. 93–109. Springer, 2024a.

Runtao Liu, I Chieh Chen, Jindong Gu, Jipeng Zhang, Renjie Pi, Qifeng Chen, Philip Torr, Ashkan Khakzar, and Fabio Pizzati. Alignguard: scalable safety alignment for text-to-image generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17024–17034. IEEE, 2025a.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025b.

Yisu Liu, Jinyang An, Wanqian Zhang, Dayan Wu, Jingzi Gu, Zheng Lin, and Weiping Wang. Disrupting diffusion: Token-level attention erasure attack against diffusion-based customization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3587–3596, 2024b.

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, 2024c.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

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

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4296–4304, 2024.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510. IEEE, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. *arXiv preprint arXiv:2302.06588*, 2023.

Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 2187–2204, 2023.

Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Investigating style similarity in diffusion models. In *European Conference on Computer Vision*, pp. 143–160. Springer, 2024.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14940–14950. IEEE, 2025.

Zhenxiong Tan, Qiaochu Xue, Xingyi Yang, Songhua Liu, and Xinchao Wang. Ominicontrol2: Efficient conditioning for diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4256–4265, 2026.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Türe. What the daam: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5644–5659, 2023.

Philipp Terhorst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Serfiq: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5651–5660, 2020.

Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 conference proceedings*, pp. 1–11, 2023.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1921–1930. IEEE, 2023.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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

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.

Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024b.

Tianyi Wang, Mengxiao Huang, Harry Cheng, Bin Ma, and Yinglong Wang. Robust identity perceptual watermark against deepfake face swapping. *arXiv preprint arXiv:2311.01357*, 2023.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

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.

Jingyao Xu, Yuetong Lu, Yandong Li, Siyang Lu, Dongdong Wang, and Xiang Wei. Perturbing attention gives you more bang for the buck: Subtle imaging perturbations that efficiently fool customized diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24534–24543. IEEE, June 2024a. doi: 10.1109/cvpr52733.2024.02316. URL http://dx.doi.org/10.1109/cvpr52733.2024.02316.

Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

Xide Xu, Muhammad Atif Butt, Sandesh Kamath, and Bogdan Raducanu. Privacy protection in personalized diffusion models via targeted cross-attention adversarial attack. *arXiv preprint arXiv:2411.16437*, 2024b.

Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18381–18391. IEEE, 2023.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824. IEEE, 2023.

Yang Zhang, Er Jin, Yanfei Dong, Yixuan Wu, Philip Torr, Ashkan Khakzar, Johannes Stegmaier, and Kenji Kawaguchi. Minimalist concept erasure in generative models. *arXiv preprint arXiv:2507.13386*, 2025.

Zhenzhu Zheng and Xi Peng. Self-guidance: improve deep neural network generalization via knowledge distillation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3451–3460. IEEE, 2022.

A Appendix

In this supplement, we provide further details on DeContext, including related work, extension experiments, implementation details, and additional analysis. Specifically, Sec. B reviews related work, Sec. C studies protection robustness under image preprocessing, Sec. D presents additional experiments on object items to evaluate the generalizability of our method to a new domain, Sec. E provides visualizations on human portraits, Sec. F presents the pseudocode, Sec. G reports user studies on preferences for protection methods, and Sec. H discusses limitations.

B Related Work

B.1 Conditional Image Generation

Diffusion models have revolutionized controllable image synthesis (Rombach et al., 2022; Podell et al., 2024; Saharia et al., 2022; Balaji et al., 2022). While early approaches rely on text prompts (Nichol et al., 2021; Ramesh et al., 2022; Ding et al., 2021), recent advances enable diverse visual conditioning through spatial structure (Zhang et al., 2023; Mou et al., 2024; Qin et al., 2023), visual features (Ye et al., 2023; Chen et al., 2024b; Yang et al., 2023; Tan et al., 2025; 2026), and style transfer (Wang et al., 2024b; Liu et al., 2023; Frenkel et al., 2024). Most recently, large-scale DiT-based models like Flux Kontext (Labs et al., 2025), Step1X-Edit (Liu et al., 2025b), and Qwen-Image (Wu et al., 2025) employ architectures that separate text-stream and image-stream processing, using reference images as primary conditioning signals. Unlike T2I approaches requiring fine-tuning (Ruiz et al., 2023; Gal et al., 2022; Kumari et al., 2023; Tewel et al., 2023), these in-context image editing models directly condition generation on context images at inference time. While this offers powerful control and ease of use, it also introduces distinct privacy risks: personal images can be easily manipulated for identity impersonation or misinformation without owner consent, making protection mechanisms critically needed.

B.2 Privacy Protection in Generative Models

Growing concerns over unauthorized image use have driven the development of adversarial protection methods. Training-time defenses disrupt model personalization (Van Le et al., 2023; Liang & Wu, 2023; Shan et al., 2023; Liang et al., 2023; Liu et al., 2024c; Wang et al., 2023), while inference-time approaches protect against editing tasks (Salman et al., 2023; Huang et al., 2021). Recent attention-aware methods (Xu et al., 2024a; Liu et al., 2024b; Xu et al., 2024b) incorporate attention objectives but treat them as auxiliary losses in optimization. However, these methods face critical limitations for modern DiT-based I2I models. First, training-time defenses (Van Le et al., 2023; Liang & Wu, 2023; Liu et al., 2024c) are ineffective against inference-only conditioning. Second, encoder or latent-targeted approaches (Salman et al., 2023; Xu et al., 2024a; Huang et al., 2021) do not account for how contextual information propagates through the dual-stream architecture. Third, existing attention-aware methods (Xu et al., 2024a;b; Liu et al., 2024b) lack systematic analysis of the specific attention mechanisms that mediate context-to-output information flow in large-scale DiT I2I frameworks. Understanding attention is therefore essential for designing effective defenses. Most closely related, NullEdit (Huang et al., 2026) redirects VLM conditioning to suppress unauthorized edits, while DeContext directly disrupts context propagation within DiT multimodal attention.

B.3 Attention Mechanisms in Diffusion Models

Transformer-based diffusion architectures rely on cross-attention to integrate multimodal inputs (Vaswani et al., 2017; Esser et al., 2024; Peebles & Xie, 2023; Bao et al., 2023; Tan et al., 2025; 2026). In dual-stream I2I models, cross-attention layers explicitly mediate context-target interactions, with attention weights determining conditioning strength. Recent interpretability studies (Tang et al., 2023; Zheng & Peng, 2022) show that attention mechanisms are both analyzable and manipulable. Critically, attention influence varies across denoising timesteps (Hertz et al., 2022; Tumanyan et al., 2023; Cao et al., 2023) and transformer depths (Karras et al., 2024), revealing that context propagates non-uniformly through the generation process. These findings suggest that targeted temporal and spatial interventions can effectively modulate conditioning pathways without global model modifications, providing a principled foundation for attention-based defenses. Our work addresses these gaps

by providing the first systematic analysis of context propagation in DiT I2I models and proposing DeContext, a defense that directly targets cross-attention at critical timesteps and blocks, achieving effective protection while maintaining output quality.

C Robustness against common purifications

metrics	Clean	None*	JPEG75	JPEG90	Gaussianσ=0.5	TVλ=0.04
ISM↓	0.79	**0.13**	0.32	0.27	0.20	0.35
CLIP-I↓	0.95	**0.49**	0.67	0.62	0.50	0.66

Table 7: Performance under purifications.

We examine the robustness of DeContext against typical pixel-level purification methods, including JPEG compression (Dziugaite et al., 2016) (quality 75 and 90), Gaussian smoothing (Xu et al., 2017) ($\sigma = 0.5$), and total variation minimization (Guo et al., 2017) ($\lambda = 0.04$). For evaluation, we sample 20 distinct identities and generate 5 images per identity. As shown in Tab. 7, context images protected by DeContext consistently yield substantially lower ISM and CLIP-I scores even after sanitization. This indicates that our defense remains effective and robust under post-processing purifications.

D Extension Experiments on Other Domains

In the main paper, we focus primarily on human face generation to reduce the risk of misuse. To further examine whether DeContext generalizes beyond the portrait domain, we extend our experiments to common objects, artistic styles, and natural scenes. Quantitative results in Tabs. 8 and 9 show that DeContext reduces context similarity across these domains beyond faces.

Extension to Items. We select 50 item images from the OminiControl Subject200K dataset (Tan et al., 2025), and resize all images to 512×512. We follow the same training settings as in the human portrait experiments, with the only modification being the prompt pool, which is adjusted to better suit item-centered content. For each clean and perturbed item image, we generate 10 outputs for every image–prompt pair using different random seeds.

Evaluation Metrics. We employ the following metrics to quantitatively evaluate the results:

- *DINO* (Caron et al., 2021): Extracts image features to evaluate semantic similarity of items, focusing on shape and structure regardless of background.
- *CLIP-Image* (Hessel et al., 2021): Uses image embeddings to measure visual-semantic consistency, capturing category and style-level similarity.
- *SSIM* (Wang et al., 2004): Computes structural similarity to assess pixel-wise fidelity between generated and reference images.

Results. We evaluate item-level context detachment using DINO, CLIP-Image, and SSIM by comparing images generated from perturbed versus clean contexts. As shown in Tab. 8, across six prompts, the average similarity scores decrease by 58%, 25%, and 64%, respectively, indicating substantial divergence and successful contextual decoupling. Visual examples in Fig. 7 further confirm that our method offers a reasonable defense against image misuse.

	"A photo of this item."			**"A top-up view of this item."**			**"Make this item look worn and old."**		
Settings	**DINO↓**	**CLIP-I↓**	**SSIM↓**	**DINO↓**	**CLIP-I↓**	**SSIM↓**	**DINO↓**	**CLIP-I↓**	**SSIM↓**
Clean	0.97	0.98	0.61	0.78	0.87	0.22	0.90	0.93	0.47
DeContext	0.36	0.65	0.24	0.32	0.66	0.14	0.28	0.59	0.09
	"Illuminate this item from left."			**"Change the color of this item to blue."**			**"Place this item onto a mirror."**		
Settings	**DINO↓**	**CLIP-I↓**	**SSIM↓**	**DINO↓**	**CLIP-I↓**	**SSIM↓**	**DINO↓**	**CLIP-I↓**	**SSIM↓**
Clean	0.98	0.97	0.85	0.94	0.94	0.79	0.87	0.91	0.70
DeContext	0.53	0.63	0.17	0.35	0.65	0.28	0.45	0.68	0.25

Table 8: Quantitative results for items under different prompts.

	Artistic-Style Protection				Scene Protection			
	"A mountain landscape."		"Add a woman."		"Add garbage."		"Snowy winter scene."	
Settings	**CSD-Sim↓**	**CLIP-I↓**	**CSD-Sim↓**	**CLIP-I↓**	**DINOv2↓**	**CLIP-I↓**	**DINOv2↓**	**CLIP-I↓**
Clean	0.14	0.54	0.73	0.74	0.59	0.83	0.51	0.75
DeContext	0.05	0.44	0.32	0.49	0.36	0.67	0.09	0.66

Table 9: Quantitative results for artistic-style and scene protection under different prompts. The prompts are abbreviated for readability.

Extension to Artistic Styles and Natural Scenes. We further evaluate DeContext on 10 artworks from five artists and 10 natural images. For artworks, we use two prompts: "Turn the scene into a mountain landscape while maintaining the artistic style" and "Add a woman to this scene while maintaining the artistic style." For natural scenes, we use "Add some garbage to this scene" and "Turn this scene into a snowy winter scene." We evaluate artistic-style similarity using *CSD-Sim* (Somepalli et al., 2024) and scene similarity using *DINOv2* (Oquab et al., 2023), and report *CLIP-Image* similarity for both domains.

As shown in Tab. 9, DeContext reduces all reported similarity scores across the four prompts. For artworks, CSD-Sim decreases from 0.14 to 0.05 for the mountain prompt and from 0.73 to 0.32 for the woman prompt. For natural scenes, DINOv2 similarity decreases from 0.59 to 0.36 for the garbage prompt and from 0.51 to 0.09 for the snowy winter prompt. These results provide further evidence that DeContext can disrupt context preservation beyond facial identity, including artistic styles and natural scenes.

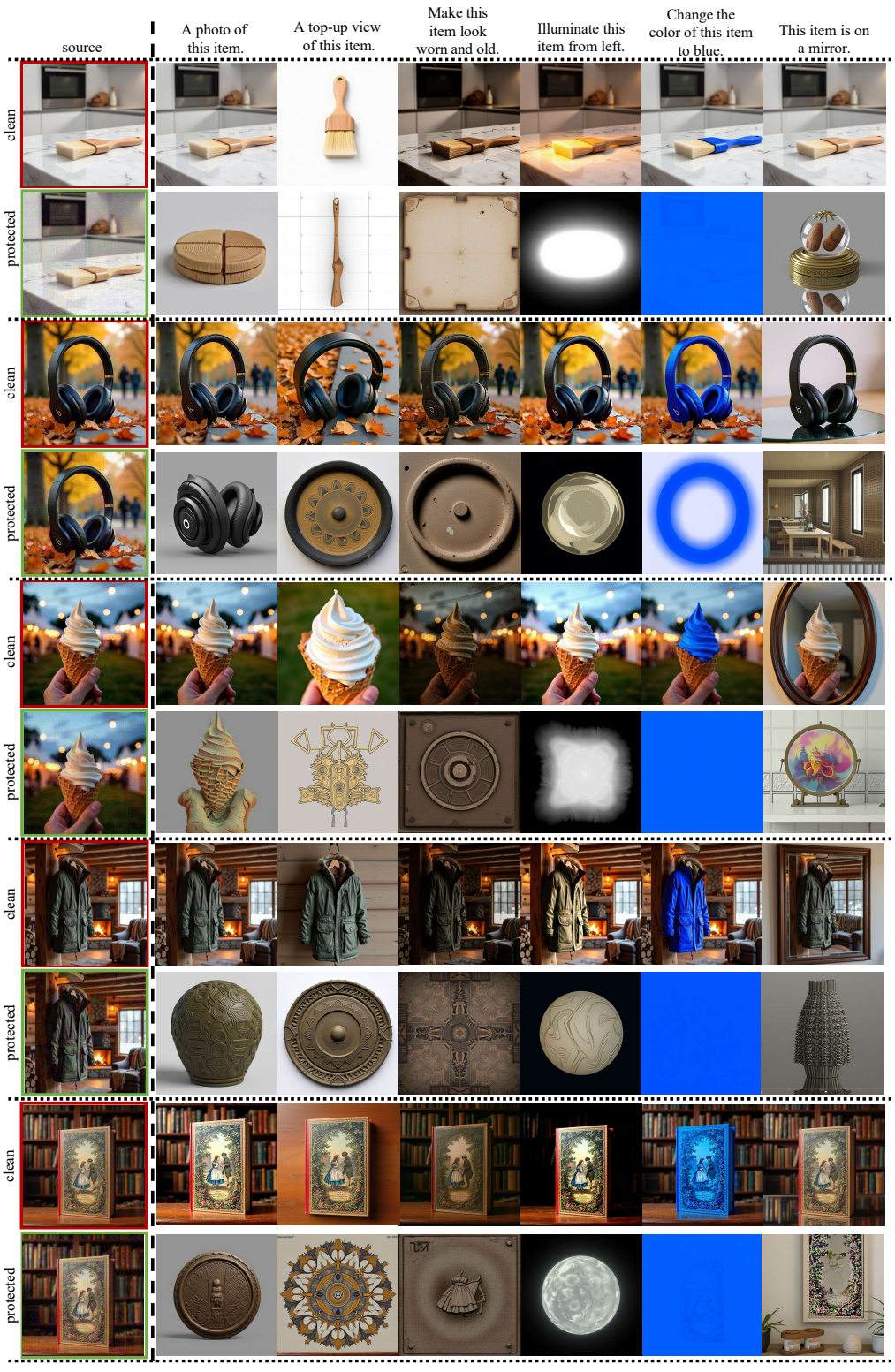


Figure 7: Defense on item images.

E ADDITIONAL QUALITATIVE RESULTS

We provide more visualization results of human portraits defense in Fig.8.

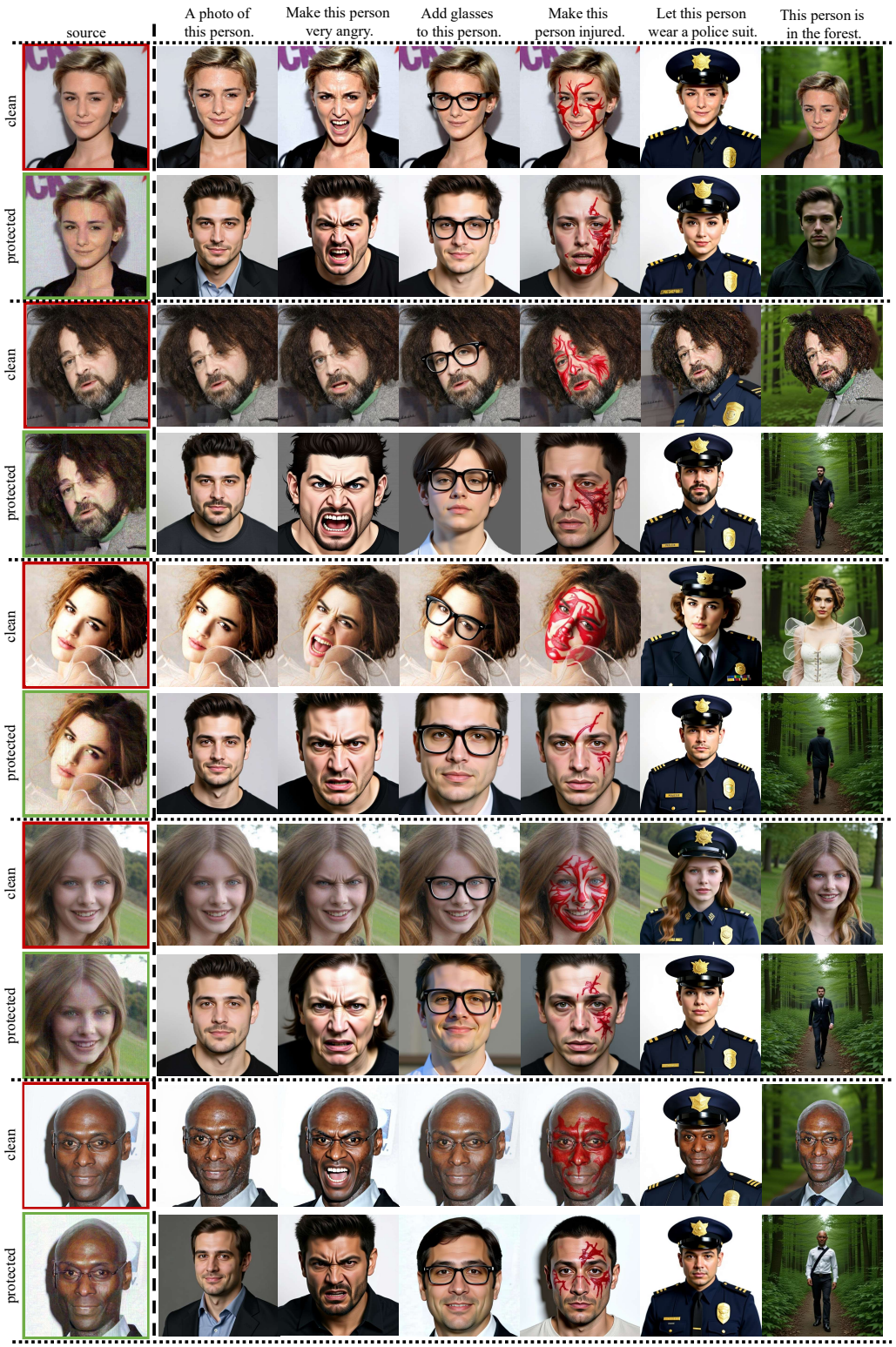


Figure 8: Defense on human portraits.

F Implementation Details

F.1 Diff-PGD Baseline

We first implement an intuitive adversarial baseline **Diff-PGD**, where perturbations on the condition image $x^{\mathrm{adv}} = x + \delta$ are optimized to maximize the *flow matching* loss. The goal is to prevent the model from generating the desired output when conditioned on x^{adv}.

Specifically, at each sampled diffusion timestep t, the I2I model $\mathcal{G}$ predicts the velocity term:

$$\hat{v}_t = \mathcal{G}\big(z_t,\ \mathcal{E}_{\mathrm{vae}}(x^{\mathrm{adv}}),\ p,\ t\big)\,,$$

where z_t denotes the noisy latent at timestep t, $\mathcal{E}_{\mathrm{vae}}(\cdot)$ encodes the perturbed condition image into latent space, and p is a randomly sampled prompt. The ground truth target latent is given by z_t^*, and the reconstruction-driven objective encourages the predicted velocity to deviate from this target to achieve protection:

$$\mathcal{L}_{\mathrm{recon}} = \mathbb{E}_{t,p,z}\Big[\|\hat{v}_t - (z_t - z_t^*)\|_2^2\Big]\,.$$

The perturbation δ is optimized via projected gradient ascent under an ℓ_∞ constraint, as shown in Algorithm 1.

Algorithm 1 Diff-PGD: Reconstruction Attack

Require: Condition image x, prompt pool $\mathcal{P}$, target latents $\{z_i^*\}$, diffusion scheduler $\mathcal{S}$, step size α, perturbation bound ϵ, total steps N.
Ensure: Perturbed image x^{adv}.

Initialize $\delta \leftarrow 0, \quad x^{\mathrm{adv}} \leftarrow x$
for $i = 1$ to N **do**
  Sample $p_i \sim \mathcal{P}$, $t \sim \mathcal{U}(t, T)$, $z \sim \mathcal{N}(0, I)$
  $\sigma_t \leftarrow \mathcal{S}(t), \quad z_t \leftarrow (1 - \sigma_t)z^* + \sigma_t n$
  $c^{\mathrm{adv}} \leftarrow \mathcal{E}_{\mathrm{vae}}(x^{\mathrm{adv}})$
  Forward through I2I diffusion model $\mathcal{G}$
  $\hat{v} \leftarrow \mathcal{G}(\tilde{z}, c^{\mathrm{adv}}, p_t, t)$
  $\mathcal{L} \leftarrow \|\hat{v} - (z_t - z^*)\|_2^2$
  $g \leftarrow \nabla_{x^{\mathrm{adv}}}\mathcal{L}$
  $x^{\mathrm{adv}} \leftarrow x^{\mathrm{adv}} + \alpha\,\mathrm{sign}(g)$
  $x^{\mathrm{adv}} \leftarrow \mathrm{clip}(x^{\mathrm{adv}},\ x - \epsilon,\ x + \epsilon)$
end for
return x^{adv}

Regardless of whether we target the early denoising stages (higher timesteps), which are critical for propagating context information, Diff-PGD fails to produce meaningful adversarial effects in DiT-based image-to-image models.

F.2 DeContext

We then implement DeContext to intentionally suppress the influence of context images by weakening their cross-attention contributions. The detailed attack strategy is introduced in Algorithm 2.

For the concentrated attack setting, we target single blocks from 0 to 25 and sample timesteps in the range of 980 to 1000 by default. Although later layers are not directly optimized, they may be affected indirectly by the disrupted signals coming from the attacked blocks. Noticed that, at the selected high-noise steps, we approximate the target image with random Gaussian noise $z \sim \mathcal{N}(0, I)$, avoiding the need for paired target images corresponding to each prompt, making the optimization both simpler and more flexible.

Algorithm 2 DeContext: Attention-aware Attack.

Require: Condition image x, prompt pool $\mathcal{P}$, diffusion scheduler $\mathcal{S}$ with T timesteps, step size α, perturbation bound ϵ, total steps N.
Ensure: Perturbed image x^{adv}.

Initialize $\delta \leftarrow 0, \quad x^{\text{adv}} \leftarrow x$
for $i = 1$ to N **do**
  Sample $p_i \sim \mathcal{P}$, $t_i \sim \mathcal{U}(t_{\text{high}}, T)$, $z \sim \mathcal{N}(0, I)$
  $z_t \leftarrow \mathcal{S}(z, t_i), \quad c^{\text{adv}} \leftarrow \mathcal{E}_{\text{vae}}(x^{\text{adv}})$
  Single blocks forward through Transformer $\mathcal{T}$
  $(q, k) \leftarrow \mathcal{T}(z_t, c^{\text{adv}}, p_i, t)$
  $A_{q,:} \leftarrow \text{Softmax}\big((qk^\top)/\sqrt{d}\big)$ ▷ target querys
  $A_{\text{ctx}} = \{A_{qk} \mid k \in \mathcal{C}\}$ ▷ context keys
  $r_{\text{ctx}} \leftarrow \frac{1}{HB|Q|} \sum_{h,b} \sum_{q \in Q} \sum_{k \in \mathcal{C}} A_{qk}^{(h,b)}$
  ▷ mean attention over heads, blocks, queries
  $\mathcal{L} \leftarrow 1 - r_{\text{ctx}}, \quad g \leftarrow \nabla_{x^{\text{adv}}} \mathcal{L}$
  $x^{\text{adv}} \leftarrow \text{clip}(x^{\text{adv}} + \alpha \cdot \text{sign}(g),\ x - \epsilon,\ x + \epsilon)$
end for
return x^{adv}

F.3 Prompt Pool

To promote prompt-agnostic generalization during adversarial optimization, we construct a diverse prompt pool $\mathcal{P}$ containing 60 natural-language instructions categorized into six groups: facial expressions, accessory addition, posture and movement, style changes, scene changes, and combined edits.

(1) Facial Expressions

- Make this person smile happily.
- Make this person look angry.
- Make this person look surprised.
- Make this person look sad.
- Give this person a serious expression.
- Make this person laugh joyfully.
- Make this person look worried.
- Make this person look exciteppd.
- Make this person look tired.
- Make this person look confused.

(2) Accessory Addition

- Add glasses to this person.
- Add sunglasses to this person.
- Add a hat to this person.
- Add a scarf to this person.
- Add a necklace to this person.
- Add earrings to this person.
- Add a cap to this person.
- Add a headband to this person.
- Add a watch to this person.
- Add a tie to this person.

(3) Posture and Movement

- Make this person wave their hand.
- Make this person cross their arms.
- Make this person point forward.
- Make this person give a thumbs up.
- Make this person put hands on hips.
- Make this person cover their mouth.
- Make this person hold up a peace sign.
- Make this person shrug their shoulders.
- Make this person touch their chin.
- Make this person salute.

(4) Style Changes

- Make this person look older.
- Make this person look younger.
- Add a beard to this person.
- Add a mustache to this person.
- Make this person's hair longer.
- Make this person's hair shorter.
- Add makeup to this person.
- Add freckles to this person.
- Make this person's skin tanned.
- Add stubble to this person.

(5) Scene Changes

- Place this person on a tropical beach with palm trees and ocean waves.
- Put this person in a snowy mountain landscape with pine trees.
- Move this person to a bustling city street with skyscrapers.
- Put this person in a beach setting.
- Put this person in a snowy mountain.
- Put this person in a city street.
- Put this person in a forest.
- Put this person in a coffee shop.
- Put this person in a library.
- Put this person in a park.

(6) Combined Prompts

- Make this person smile and wave.
- Add glasses and make this person look serious.
- Make this person look older with a beard.
- Add a hat and make this person point forward.
- Make this person look surprised with hands on face.
- Add sunglasses and make this person give a thumbs up.

- Make this person look younger and smile happily.
- Add a scarf and make this person cross their arms.
- Make this person look tired and rub their eyes.
- Add and make this person look confident.

G User Study

To evaluate the subjective effectiveness of our defense, we conducted a controlled user study with 20 participants. We compared **DeContext** with four state-of-the-art image protection methods: Anti-Dreambooth (Van Le et al., 2023), CAAT (Xu et al., 2024a), AdvDM (Liang & Wu, 2023), and FaceLock (Wang et al., 2025).

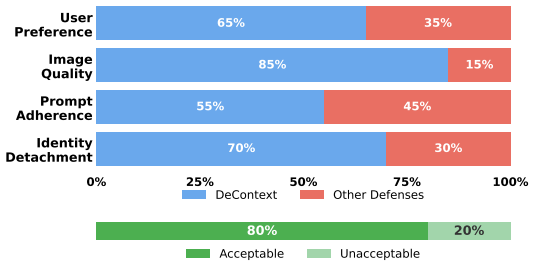


Figure 9: User Study.

Each participant viewed 8 generated images per method using the same prompt and protected image. They ranked results along four dimensions: (1) *Identity Detachment*, (2) *Prompt Adherence*, (3) *Image Quality*, and (4) *Overall Protection Preference*. We also assessed *User Acceptability* of the perturbed contexts produced by DeContext.

As shown in Fig. 9, DeContext obtained the most top-ranked selections across all dimensions. Its total number of first-place rankings exceeded the combined first-place counts of the four baselines, indicating strong and consistent user preference for DeContext. Additionally, supporting imperceptibility, 80% of participants rated the perturbed inputs visually acceptable for everyday use.

H Discussion and Future Work

Failure Case on Complex Scenes. DeContext works well in simple image edits where the context image plays a dominant role. However, its effectiveness diminishes under complex scene modifications driven by language.

As shown in Fig. 10, when the prompt demands a substantial transformation of the scene, the model largely disregards the original visual context, even in the absence of any defense mechanism. In such cases, generation is primarily governed by the text prompt rather than the input image. Because the model already allocates minimal attention to the visual context, further weakening cross-attention via DeContext yields only minor effects (e.g., slight resizing of a person), leaving object identity and overall scene composition largely intact. This demonstrates that strong textual guidance can override contextual signals, thereby limiting the impact of uniform context detachment.

Future Direction. Given these limitations, a promising avenue for future work is to shift from uniform suppression of contextual information toward selective attention reduction. By targeting the reduction more precisely at sensitive elements, the method may better limit their influence, especially in scenarios where the prompt requires significant scene alterations.

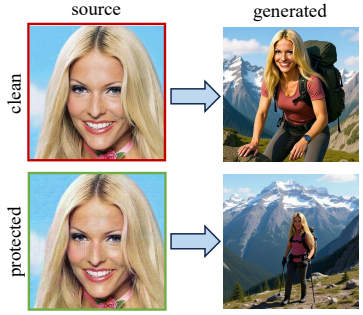


Figure 10: Failure Case with the prompt "Transform this image to mountain hiking scene, change the person's pose to climbing stance, add backpack and snow peaks."