TRANSFERABLE VISUAL IDENTITY PROTECTION AGAINST UNAUTHORIZED REUSE IN CLOSED-SOURCE IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Image-conditioned generation raises concerns about unauthorized use of visual identity. Adversarial image protection offers a lightweight defense against such misuse with small perturbations. However, existing defenses are typically developed on accessible generative pipelines, and their transfer across modern closed-source editors remains insufficiently characterized. To this end, we propose T-VIP, a transferable visual identity protection method for closed-source models that embeds a misleading identity signal into the protected image. Specifically, T-VIP constructs a transferable identity representation space from an ensemble of image encoders with lightweight projectors amplifying identity-discriminative cues. It then optimizes a push-and-pull misleading objective that drives the protected representation away from the source identity while keeping it within an identity-relevant manifold. Extensive experiments demonstrate that T-VIP substantially outperforms existing protection methods across diverse image generators without requiring white-box access. More importantly, it effectively transfers to closed-source systems, including GPT Image 2, Nano Banana 2, and Grok Imagine, highlighting its practicality for real-world visual identity protection.

1 INTRODUCTION

The use of reference images in visual generation has evolved from specialized editing and personalization mechanisms toward increasingly general multimodal control. Early diffusion systems enabled image-guided synthesis and instruction-guided editing (Meng et al., 2022; Brooks et al., 2023), while personalization allowed specific subjects to be reproduced from a small reference set (Ruiz et al., 2023). Modern multimodal generators jointly follow visual references and natural-language instructions, preserving identity while changing pose, clothing, style, or scene (Wu et al., 2025; 2026). Commercial services such as Nano Banana 2 and GPT Image 2 have made these capabilities broadly accessible (Google DeepMind, 2026; OpenAI, 2026).

Despite these advances, image-conditioned generation introduces a serious risk of unauthorized identity reuse. A publicly released personal image can be used as a reference to synthesize identity-preserving variants without the subject's consent. As shown in Figure 1, given only a clean portrait, commercial closed-source generators can readily produce realistic, identity-preserving variants. Large-scale user evidence suggests that people distinguish AI-generated images from real photographs only modestly above chance (Roca et al., 2025; Li et al., 2025a).

Adversarial image protection (Zhu et al., 2023) offers a lightweight defense by adding visually small perturbations to images before release, aiming to disrupt a generator's ability to preserve and reuse the protected subject. Anti-DreamBooth targets DreamBooth-style personalization (Van Le et al., 2023). PhotoGuard and AdvPaint perturb images to disrupt diffusion-based image-to-image editing or masked inpainting pipelines (Salman et al., 2023; Jeon et al., 2025). DeContext further studies protection against in-context DiT editing (Shen et al., 2025). Prior work has reported transfer to a small number of proprietary systems, but evaluation across a broad suite of current commercial reference-conditioned editors under one unified protocol remains limited (Song et al., 2024; 2025).

To address this gap, we propose T-VIP, a visual identity protection method for closed-source models that embeds a transferable misleading identity signal into the protected image. As illustrated in

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: **Black-box Identity Protection.** Nano Banana Pro ✦, GPT Image 2 ֍, and Grok Imagine ∅, preserve the identity from clean images. T-VIP redirects the same edits to a mismatched identity, preventing unauthorized reuse. Images generated from T-VIP-protected references show consistent changes in hairstyle, gender presentation, and age.

Figure 2, a counterfactual misleading prompt shows that subject preservation can be redirected by conflicting identity cues. T-VIP implements this redirection at the image level, so that black-box generators preserve the wrong identity under ordinary editing prompts.

Specifically, T-VIP is built around two components. First, it constructs a transferable identity representation space from an ensemble of image encoders, using lightweight projectors to amplify identity-discriminative cues. Second, it optimizes a misleading objective that drives the protected representation away from the source identity while keeping it within an identity-relevant manifold. Under a small perturbation budget, T-VIP produces visually faithful images whose identity representation is shifted toward a different yet plausible identity while retaining high perceptual fidelity.

Extensive experiments demonstrate that T-VIP substantially improves visual identity protection across modern image-conditioned generators. On Qwen-Image-Edit-2511, it reduces Visual Identity Similarity (VIS) to 0.184 without target-model access. The resulting protection also transfers directly to commercial closed-source systems, yielding a VIS of 0.082 on Nano Banana 2, where the evaluated prior defenses remain above 0.7.

Overall, our contributions are: (i) We formulate identity protection for commercial closed-source generators as a black-box problem. (ii) We propose T-VIP, combining a transferable identity space with a misleading objective. (iii) Extensive experiments demonstrate its effectiveness across diverse open-weight and closed-source generators.

2 Related Work

Image-Conditioned Generation. Image-conditioned generation has evolved from subject-specific personalization methods such as DreamBooth (Ruiz et al., 2023) to general image-input editing systems such as FLUX.1 Kontext (Black Forest Labs et al., 2025). Their conditioning designs vary. Step1X-Edit uses an MLLM to encode the reference image and editing instruction before diffusion decoding (Liu et al., 2025), while OmniGen2 frames generation and editing as unified multimodal generation with VLM-conditioned diffusion decoding (Wu et al., 2026). Qwen-Image separately encodes semantic and reconstructive representations through a vision–language model and a VAE (Wu et al., 2025). Commercial services such as Nano Banana 2 and GPT Image 2 expose a similar image-input interface (Google DeepMind, 2026; OpenAI, 2026), but their internal conditioning architectures are not disclosed in the cited product documentation.

Adversarial Image Protection. Adversarial image protection optimizes visually small perturbations on released images so that downstream generative models cannot faithfully reuse the protected visual identity. Earlier protections largely assume personalization or customization misuse, where released images are used to fine-tune or adapt a generator. Anti-DreamBooth targets DreamBooth-style fine-tuning, while ACE shows that targeted objectives strengthen protection across customization routes (Van Le et al., 2023; Zheng et al., 2025). PhotoGuard and EditShield study image-to-image or instruction-guided editors (Salman et al., 2023; Chen et al., 2024). DiffusionGuard and AdvPaint focus on masked editing and inpainting (Choi et al., 2025; Jeon et al., 2025), while DeContext studies in-context DiT editing (Shen et al., 2025). DiffVax and UniDef further improve efficiency and transferability for diffusion editing protection (Ozden et al., 2026; Shao et al., 2026). IDProtector learns a feed-forward adversarial noise encoder for efficient protection against encoder-based



Figure 2: **Intuition of T-VIP.** We probe Nano Banana Pro with an ordinary editing prompt and a misleading prompt that asserts a counterfactual attribute. (*Left*) The counterfactual prompt shows that subject preservation is sensitive to conflicting identity cues, steering the edit toward a mismatched identity. (*Right*) T-VIP embeds transferable misleading identity cues into the image itself, leading black-box generators to preserve the wrong identity under an ordinary editing prompt.

generation (Song et al., 2025). IDGuardian further disrupts both identity extraction and injection for training-free and training-based personalization (Xiong et al., 2026). However, their real-world deployability remains contested, as imperceptible protections can be weakened by common transformations and model variation (Zhao et al., 2024; Hönig et al., 2025).

Differences with Prior Work. We systematically study transferable visual identity protection across modern open-weight and closed-source image generators. Methodologically, T-VIP prioritizes cross-model transferability, whereas prior defenses typically formulate their objectives around particular accessible models or pipelines.

3 Method

3.1 Formulation

Problem Statement. Given a source image x^{src}, we seek a protected counterpart x^{prt} that remains visually faithful to the source while preventing downstream image-conditioned generators from reusing the source identity. We consider a black-box setting, where the downstream prompt and the generator's weights are unavailable at protection time. Formally, an unseen image-conditioned generator G produces $x^{\mathrm{gen}} = G(x^{\mathrm{prt}}, c)$ under an unknown instruction c, and the desired protected image satisfies:

$$\mathrm{ID}(x^{\mathrm{gen}}) \neq \mathrm{ID}(x^{\mathrm{src}}), \quad \left\|x^{\mathrm{prt}} - x^{\mathrm{src}}\right\|_p \leq \epsilon, \tag{1}$$

where $\mathrm{ID}(\cdot)$ denotes the visual identity conveyed by an image, and ϵ specifies a standard ℓ_p perturbation budget that limits visible distortion.

T-VIP Overview. T-VIP addresses this problem by embedding a transferable misleading identity signal into the protected image. As shown in Figure 3, its design comprises two components. First, T-VIP constructs an identity encoder that captures transferable identity cues:

$$\Phi : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{d}. \tag{2}$$

Here, transferability means that the identity cues captured by Φ are likely to remain relevant to the conditioning pathways of inaccessible generators G.

Second, motivated by the intuition in Figure 2, T-VIP embeds a misleading identity signal into the protected image through optimization in this identity space:

$$\min_{x^{\mathrm{prt}}} \mathcal{L}_{\mathrm{mislead}}\left(x^{\mathrm{prt}}, x^{\mathrm{src}}\right) \quad \text{s.t. } \left\|x^{\mathrm{prt}} - x^{\mathrm{src}}\right\|_p \leq \epsilon, \tag{3}$$

The following sections instantiate the identity encoder $\Phi(\cdot)$ and the misleading objective $\mathcal{L}_{\mathrm{mislead}}$.

3.2 Transferable Identity Space

To model identity structure shared across diverse architectures, we ensemble pretrained vision encoders and enhance their identity-discriminative representations with lightweight projection heads.

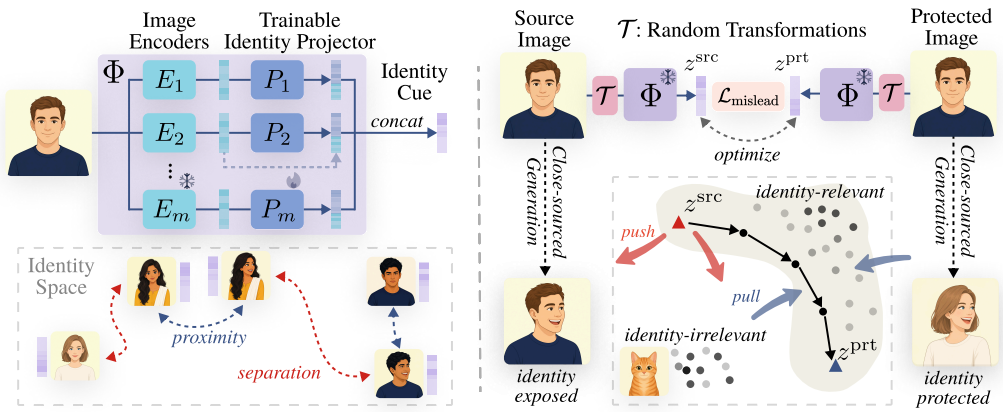

Figure 3: **Overview of T-VIP.** (*Left*) T-VIP constructs a transferable identity space from an ensemble of frozen image encoders. Lightweight projectors are optimized with contrastive supervision to amplify identity-discriminative cues. (*Right*) T-VIP then steers the protected representation away from the source identity and toward a plausible misleading representation within the identity-relevant manifold. This misleading identity cue protects the source identity in closed-source generation.

Specifically, the surrogate identity encoder Φ ensembles M vision encoders. Combining multiple backbones reduces reliance on encoder-specific features and improves transferability to unseen architectures (Liu et al., 2017). For an input image x, let h_m denote the feature produced by the m-th ensemble member. We concatenate these features and normalize:

$$\Phi(x) = \frac{[h_1, \ldots, h_M]}{\|[h_1, \ldots, h_M]\|_2}, \tag{4}$$

where $[\cdot]$ denotes feature concatenation. Each ensemble member concatenates a normalized frozen encoder feature e_m and a projected identity feature r_m:

$$\begin{aligned} h_m &= \frac{[e_m, r_m]}{\|[e_m, r_m]\|_2}, \\ e_m = \frac{E_m(x)}{\|E_m(x)\|_2}, &\qquad r_m = \frac{P_m e_m}{\|P_m e_m\|_2}. \end{aligned} \tag{5}$$

Here, E_m is the frozen encoder, and P_m is a lightweight identity projection head. During construction, only P_m is trained on an auxiliary identity-labeled image set. We use a supervised contrastive objective over mini-batches, pulling projected features of the same identity together while contrasting them against other images in the batch. Consequently, each h_m preserves the encoder's general visual representation while placing emphasis on identity-discriminative cues.

3.3 Misleading Identity Signal

As motivated in Figure 2, we aim for the misleading signal to diverge from the source while remaining anchored to a plausible identity. We realize these two requirements with a push-and-pull misleading objective:

$$\mathcal{L}_{\text{mislead}}\left(x^{\text{prt}}, x^{\text{src}}\right) = \mathcal{L}_{\text{push}} + \alpha \mathcal{L}_{\text{pull}}, \tag{6}$$

where the push term separates the protected representation from the source identity, the pull term anchors it to plausible alternative identities, and α balances the two terms.

Both terms are defined in the transferable identity space above. To reduce overfitting to a fixed input view during optimization, we average the representation over differentiable random transformations (Xie et al., 2019):

$$z(x) = \mathbb{E}_{\mathcal{T}\sim\mathcal{D}}\left[\Phi(\mathcal{T}(x))\right], \tag{7}$$

where $\mathcal{D}$ is a distribution of random resized crops. Accordingly, $z^{\text{prt}} = z(x^{\text{prt}})$ and $z^{\text{src}} = z(x^{\text{src}})$.

The push term reduces source identity similarity:

$$\mathcal{L}_{\text{push}}\left(x^{\text{prt}}, x^{\text{src}}\right) = \cos\left(z^{\text{prt}}, z^{\text{src}}\right). \tag{8}$$

Minimizing this term discourages the protected representation from preserving the source identity.

The pull term attracts the protected representation toward an auxiliary bank of plausible alternative identities, $\mathcal{A} = \{z_i^{\text{alt}}\}_{i=1}^{B}$, where z_i^{alt} is the representation of an auxiliary image x_i^{alt} in the transferable identity space. We use a soft nearest-neighbor objective:

$$\mathcal{L}_{\text{pull}}\left(x^{\text{prt}}, \mathcal{A}\right) = -\tau \log \frac{1}{B} \sum_{i=1}^{B} \exp\left(\frac{s_i}{\tau}\right), \qquad s_i = \cos\left(z^{\text{prt}}, z_i^{\text{alt}}\right), \tag{9}$$

where τ is a temperature.

Finally, we solve the constrained optimization of $\mathcal{L}_{\text{mislead}}$ in Eq. (3) using an iterative update-and-projection loop for N steps (Madry et al., 2018), illustrated below with a basic gradient step:

$$g = \nabla_{x^{\text{prt}}} \mathcal{L}_{\text{mislead}}\left(x^{\text{prt}}, x^{\text{src}}\right), \qquad x^{\text{prt}} \leftarrow \Pi_{\mathcal{C}}\left[x^{\text{prt}} - \eta g\right]. \tag{10}$$

Here, η is the learning rate. The operator $\Pi_{\mathcal{C}}$ projects the updated image onto the feasible set $\mathcal{C} = \{x \in [0,1]^{H \times W \times C} : \|x - x^{\text{src}}\|_\infty \leq \epsilon\}$. We use Adam in practice and provide the complete algorithm in Appendix B.2.

4 Experiments

4.1 Evaluation Protocol

Given a source image x^{src}, its protected version x^{prt}, and the resulting generation $x^{\text{gen}} = G(x^{\text{prt}}, c)$, we evaluate protection effectiveness using two complementary metrics: Identity Score Matching (ISM) for measuring identity leakage in the feature space and Visual Identity Similarity (VIS) for assessing visible identity reuse.

ISM. Identity Score Matching (ISM) is a common automatic metric for measuring identity leakage in adversarial identity protection (Van Le et al., 2023; Zheng et al., 2025; Song et al., 2025). Following prior evaluations, we report ArcFace-based ISM, computed as the cosine similarity between identity embeddings of the source x^{src} and generated images x^{gen} (Van Le et al., 2023; Song et al., 2025; Deng et al., 2022). Lower ISM indicates stronger protection.

Nevertheless, ISM is sensitive to identity-irrelevant variations. Even when the visible identity is preserved, changes in head pose, facial expression, or lighting can alter face-recognition embeddings and lower their cosine similarity (Zhu et al., 2015; Liang et al., 2023). Consequently, a low ISM may reflect nuisance variation rather than genuine suppression of identity reuse.

VIS. As a complementary metric, we report Visual Identity Similarity (VIS), an MLLM-based measure of visible identity reuse. We instantiate VIS with the locally deployed Qwen3.6-27B checkpoint (Qwen Team, 2026). To reduce sensitivity to protective perturbations, we use an empirical stabilization procedure inspired by Feature-Space Smoothing (Xia et al., 2026). For each evaluation, the judge compares a contact sheet of clean reference images against the generated image and outputs a continuous identity-similarity score in $[0, 1]$. The prompt instructs the judge to assess facial identity while disregarding pose, clothing, background, lighting, and style. We report the mean continuous VIS score, with lower values indicating weaker visible identity reuse. The complete prompt, scoring protocol, design rationale, and empirical validation of VIS are provided in Appendix C.

Appendix C.1 validates VIS against human judgments. VIS outperforms ISM in correlation with human ratings (Spearman's ρ: 0.637 vs. 0.576 in Appendix Table 6) and demonstrates strong robustness to protective perturbations. We also compare results across different MLLM judges to further assess the protection effectiveness.

4.2 Experimental Setup

Datasets. Following prior facial identity protection evaluations (Van Le et al., 2023; Yang et al., 2025), we use VGGFace2 (Cao et al., 2018) as the main benchmark. Our evaluation split contains

Table 1: **Protection Effectiveness on Open-weight Models.** Results are reported on VGGFace2. Lower ISM and VIS indicate stronger protection. Best and second-best results are bolded and underlined, respectively.

Method	Target Access	Time Cost	Qwen-Edit		Step1X		LongCat		FireRed		OmniGen2		FLUX K.	
			ISM↓	VIS↓	ISM↓	VIS↓	ISM↓	VIS↓	ISM↓	VIS↓	ISM↓	VIS↓	ISM↓	VIS↓
No protection	–	–	0.305	0.795	0.562	0.852	0.345	0.768	0.321	0.792	0.392	0.736	0.327	0.677
IDProtector	Black-box	**1.2s**	0.265	0.758	0.513	0.797	0.330	0.761	0.278	0.699	0.285	0.689	0.272	0.623
IDGuardian	Black-box	35.1s	0.270	0.745	0.497	0.773	0.304	0.750	0.264	0.688	0.268	0.644	0.251	0.620
DeContext	White-box	592.4s	0.261	0.731	**0.043**	**0.190**	0.338	0.699	0.319	0.768	0.166	0.393	0.153	0.508
T-VIP (Ours)	Black-box	33.9s	**0.095**	**0.184**	0.174	0.381	**0.077**	**0.093**	**0.127**	**0.289**	**0.114**	**0.316**	**0.127**	**0.409**

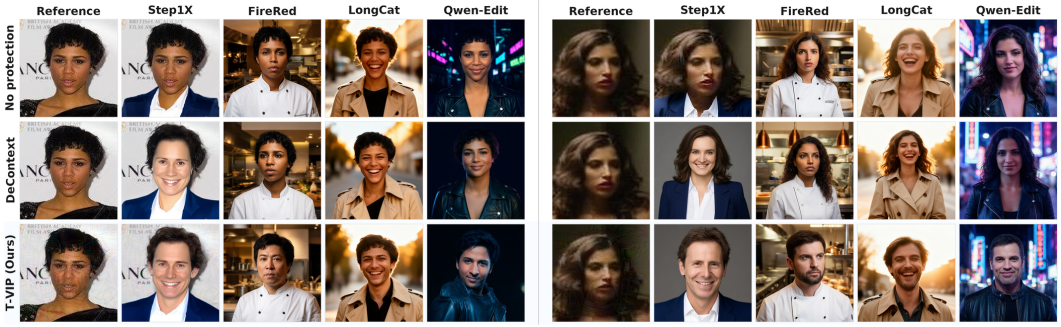

Figure 4: **Qualitative Comparison on Open-weight Models.** We present the two most effective protection methods. DeContext requires a separately optimized perturbation for each target generator. We show one such DeContext perturbation for illustration. In contrast, a single T-VIP perturbation consistently redirects all generators toward a different identity.

200 identities, and we select one source image per identity for protection evaluation. The projection heads are trained on a disjoint VGGFace2 split with no identity overlap with the evaluation set. We additionally report results on CelebA-HQ (Karras et al., 2018) in Appendix D to evaluate generalization to a higher-quality celebrity-face distribution.

Target Models. We evaluate T-VIP on six locally hosted open-weight generators: Qwen-Image-Edit-2511 (Qwen Team, 2025), Step1X-Edit (Liu et al., 2025), LongCat-Image-Edit (Meituan LongCat Team et al., 2025), FireRed-Image-Edit-1.1 (Super Intelligence Team et al., 2026), OmniGen2 (Wu et al., 2026), and FLUX.1 Kontext (Black Forest Labs et al., 2025), and six commercial closed-source systems: GPT Image 2 (OpenAI, 2026), Gemini 3.1 Flash Image (Nano Banana 2) (Google DeepMind, 2026), Grok Imagine-quality (xAI, 2026), Seedream 4.5 (BytePlus, 2025), MAI-Image-2.5 (Microsoft AI Superintelligence Team, 2026), and Riverflow 2.5 Fast (Sourceful, 2026). We use ten editing prompts, with details provided in Appendix C.2.

Baseline Selection. We compare T-VIP with three adversarial protection baselines applicable to our target models. DeContext is included only for open-weight models, where its white-box access is available (Shen et al., 2025). IDProtector and IDGuardian are evaluated as transfer-based identity-protection baselines without access to the target models (Song et al., 2025; Xiong et al., 2026). All methods are evaluated under the same target generators, prompts, and metrics. We also evaluate the complementarity of T-VIP with Anti-DreamBooth (Van Le et al., 2023). UniDef (Shao et al., 2026) and Anti-Reference (Song et al., 2024) report results on earlier diffusion pipelines, without evaluating the modern multimodal editors studied here. Appendix C details their evaluation scope.

Implementation. We use an ensemble of four image encoders: OpenAI CLIP ViT-B/16, OpenAI CLIP ViT-B/32, LAION/OpenCLIP ViT-B/32, and DINOv2 ViT-B/14 (Radford et al., 2021; Cherti et al., 2023; Oquab et al., 2024). For each frozen encoder, we train a linear 512-dimensional projection head on 8,000 VGGFace2 images from 2,000 identities, with four images per identity. The auxiliary bank is constructed from identities disjoint from both the projector-training and evalua-

Table 2: **Protection Effectiveness on Closed-source Models.** Results are reported on VGGFace2. Lower ISM and VIS indicate stronger protection. Best results are bolded.

Method	Target Access	Time Cost	GPT Image		Nano Banana		Grok Imagine		Seedream		MAI-Image		Riverflow	
			ISM↓	VIS↓	ISM↓	VIS↓	ISM↓	VIS↓	ISM↓	VIS↓	ISM↓	VIS↓	ISM↓	VIS↓
No protection	–	–	0.534	0.819	0.517	0.907	0.469	0.855	0.560	0.798	0.482	0.800	0.365	0.731
IDProtector	Black-box	**1.2s**	0.500	0.789	0.499	0.892	0.407	0.847	0.546	0.750	0.476	0.783	0.342	0.701
IDGuardian	Black-box	35.1s	0.487	0.754	0.478	0.841	0.428	0.855	0.536	0.739	0.455	0.733	0.331	0.672
T-VIP (Ours)	Black-box	33.9s	**0.065**	**0.009**	**0.121**	**0.082**	**0.094**	**0.033**	**0.308**	**0.323**	**0.084**	**0.070**	**0.070**	**0.137**



Figure 5: **Protection Visualization on Closed-source Models.** Within each panel, each row shows different models editing the same reference image using the same instruction. T-VIP consistently protects visual identity.

tion splits, preventing train–test identity leakage. The auxiliary identity bank contains $B = 200$ identities. We set the pull weight to $\alpha = 1.5$ and its temperature to $\tau = 0.04$, and optimize the perturbation for $N = 300$ steps using Adam with a learning rate of 0.005 and random-resized-crop input transformation. For fair comparison with the baselines, we use a default ℓ_∞ perturbation budget of $16/255$ to allow all baselines to operate effectively. We also conduct an ablation study of T-VIP at smaller perturbation budgets. All local experiments are conducted on eight NVIDIA RTX 4090 GPUs with a fixed random seed of 2026. More details are provided in Appendix C.5.

4.3 Main Results

Effectiveness on Open-weight Models. We compare T-VIP with three representative protection methods on modern open-weight generators in Table 1. IDProtector and IDGuardian provide limited protection on these targets, indicating limited transfer to the evaluated editors under our protocol. DeContext remains effective on some targets, such as Step1X and FLUX K., but relies on costly per-model optimization with white-box access. In contrast, T-VIP substantially reduces identity leakage across all evaluated generators under black-box access, while requiring far less optimization time than DeContext. The identity cues modeled by T-VIP transfer across MLLM-conditioned editors, diffusion-transformer editors, and unified generation models.

We further visualize the two strongest protection methods in Figure 4. DeContext performs well on Step1X, but provides noticeably weaker protection on other models. In contrast, T-VIP consistently disrupts identity preservation across the evaluated generators, redirecting their outputs toward a distinct identity through coherent changes in visual attributes such as gender presentation and hairstyle.

Effectiveness on Closed-source Models. Table 2 presents results on six commercial closed-source generators. Overall, these systems preserve source identity more strongly than their open-weight counterparts in the absence of protection, underscoring the practical importance of identity protection for widely deployed services. IDProtector and IDGuardian provide only limited protec-

Table 3: **Robustness against Post-Processing.** Each operation is applied to the T-VIP-protected reference image.

Input	Qwen-Edit		GPT Image		Nano Banana	
	ISM↓	VIS↓	ISM↓	VIS↓	ISM↓	VIS↓
No protection	0.305	0.795	0.534	0.819	0.517	0.907
T-VIP (Ours)	0.095	0.184	0.065	0.009	0.121	0.082
+ JPEG	0.101	0.215	0.102	0.009	0.123	0.141
+ Gaussian blur	0.128	0.301	0.180	0.127	0.225	0.366
+ Gaussian noise	0.143	0.359	0.147	0.067	0.133	0.146
+ IMPRESS	0.126	0.292	0.132	0.079	0.164	0.137

Table 4: **Complementarity to Prior Protection.** (Anti-DB: Anti-DreamBooth; ↓ indicates stronger protection).

Protection	DreamBooth		Qwen-Edit		Nano Banana	
	ISM↓	VIS↓	ISM↓	VIS↓	ISM↓	VIS↓
No protection	0.310	0.650	0.305	0.795	0.517	0.907
Anti-DB	0.138	0.195	0.281	0.718	0.501	0.866
T-VIP (Ours)	0.254	0.450	0.095	0.184	0.121	0.082
T-VIP + Anti-DB	0.128	0.173	0.082	0.284	0.120	0.088

tion in this setting, indicating that their perturbations transfer poorly to frontier commercial generators. In contrast, T-VIP consistently suppresses identity leakage across all evaluated services. Although their internal architectures are undisclosed, the observed transfer suggests that modern image-conditioned generators rely on identity-relevant visual cues shared with the diverse pretrained encoders used by T-VIP.

Figure 5 further shows that commercial generators remain coherent and instruction-aligned under T-VIP while producing identities that diverge markedly from the source. Across independently developed services, this identity divergence consistently appears through changes in facial appearance, gender presentation, and hairstyle. In the supplementary material, we provide a comparison video demonstrating T-VIP protection applied through the consumer-facing web interfaces.

Cross-dataset Generalization. Without retraining the identity projection heads or changing the T-VIP configuration, we further evaluate on CelebA-HQ. Across six open-weight and six closed-source generators, T-VIP reduces mean VIS from 0.889 to 0.218 and from 0.925 to 0.298, respectively (Appendix Tables 11 and 12). For open-weight targets, the mean paired identity-level VIS reduction relative to DeContext is 0.366 after averaging over ten prompts and six targets per identity (Appendix Table 13). These consistent reductions further validate the effectiveness of T-VIP.

Robustness against Post-Processing. We apply JPEG compression ($q = 85$), Gaussian blur (5×5, $\sigma = 1.5$), and Gaussian noise (variance 100 on a $[0, 255]$ scale) to the protected reference images before generation. We evaluate IMPRESS (Cao et al., 2023) using the default hyperparameters for image editing reported in the original paper. Table 3 reports results on representative open-weight and closed-source generators. T-VIP remains effective under all evaluated operations. Although blur and noise partially weaken protection, both ISM and VIS remain substantially below the unprotected baselines. Overall, these findings indicate that T-VIP is likely to survive routine image-sharing pipelines, including various forms of platform-induced compression.

Complementarity to Prior Protection. We further combine T-VIP with Anti-DreamBooth (Anti-DB) to test whether their complementary objectives can protect both personalization and frontier image-conditioned generation. For the combined defense, we first apply T-VIP and then optimize Anti-DB on the protected image. Both stages project the image onto the same ℓ_∞ ball of radius $\epsilon = 16/255$ centered at the original clean image. As shown in Table 4, Anti-DB is effective against its intended DreamBooth personalization pipeline but transfers little to Qwen-Edit or Nano Banana 2. Conversely, T-VIP strongly protects the frontier image-conditioned generators but is less effective against DreamBooth, which it was not designed to target. Combining the two yields the strongest DreamBooth protection while preserving strong protection on Qwen-Edit and Nano Banana 2. This behavior mirrors the post-processing robustness in Table 3, as the additional Anti-DB optimization largely preserves T-VIP's effectiveness on modern generators. T-VIP can therefore be augmented with pipeline-specific defenses when training-based personalization is also in scope.

4.4 Empirical Analysis

Protection-Fidelity Trade-off. Figure 6 characterizes the trade-off between protection strength and image fidelity. Increasing ϵ consistently lowers VIS but also reduces MS-SSIM, reflecting the



Figure 6: **Protection-Fidelity Trade-off.** (*Left*) VIS and MS-SSIM quantify protection and fidelity across perturbation budgets. (*Right*) Nano Banana 2 outputs show a noticeable identity shift at small budgets and an entirely different person at larger ones.

Table 5: **Component Analysis.** We ablate key design components of T-VIP. Lower is better.

Configuration	Qwen-Edit		GPT Image		Nano Banana	
	ISM↓	VIS↓	ISM↓	VIS↓	ISM↓	VIS↓
No protection	0.305	0.795	0.534	0.819	0.517	0.907
T-VIP (Ours)	0.095	0.184	0.065	0.009	0.121	0.082
Transferable identity space						
w/o Encoder Ensemble	0.196	0.399	0.232	0.330	0.264	0.392
w/o Identity Projection	0.114	0.228	0.105	0.079	0.183	0.149
Misleading identity signal						
w/o Push	0.215	0.664	0.351	0.602	0.348	0.625
w/o Pull	0.104	0.215	0.105	0.055	0.153	0.123

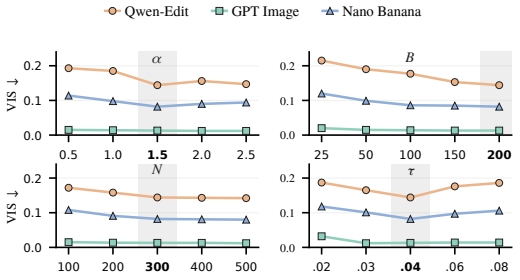

Figure 7: **Hyperparameter Sensitivity.** Default settings are shaded. Lower is better.

unavoidable modification of the source image. Notably, $\epsilon = 4/255$ already reduces the average VIS across the three generators to nearly half of its unprotected value while retaining an MS-SSIM of 0.989. The qualitative examples show the same progression. Even small perturbation budgets induce a noticeable identity shift, whereas $\epsilon = 16/255$ redirects the generation to an entirely different identity. These results make the protection level tunable, allowing the perturbation budget to be selected according to the balance between visual fidelity and identity suppression.

Component Analysis. Table 5 isolates the contribution of each component. Removing the push term causes the most significant performance degradation across all targets, establishing this formulation as the fundamental basis of our approach. Restricting the architecture to a single CLIP ViT-B/16 encoder yields a substantial drop in protection capabilities. Additionally, ablating the identity projection or the pull term consistently degrades overall performance. These results support the contributions of the encoder ensemble, identity projection, and pull objective to T-VIP's transferability.

Hyperparameter Sensitivity. Figure 7 shows that T-VIP is not highly sensitive to α or τ. VIS varies only modestly across the tested ranges, with no abrupt degradation around the defaults. Increasing B steadily improves protection, but the gains begin to saturate as the size approaches 200. We therefore set $B = 200$, obtaining a sufficiently diverse identity bank without unnecessary storage or optimization cost. VIS gradually converges as the number of optimization steps N increases; we therefore set $N = 300$ to balance protection strength and optimization time.

5 Conclusion

We presented T-VIP, a black-box protection method against unauthorized identity reuse in image-conditioned generation. T-VIP constructs a transferable identity space by projecting features from an ensemble of image encoders. Within this space, a push-and-pull objective separates the protected representation from the source identity and anchors it to plausible alternative identities, embedding a misleading identity signal while preserving visual fidelity. Experiments across diverse open-weight and commercial closed-source generators demonstrate significantly stronger and more transferable protection than prior defenses. This transferability allows an image to be protected before the downstream prompt or target generator is known. Combined with defenses for training-based personalization, T-VIP extends protection across different reuse pipelines. We hope this work advances practical identity protection for the rapidly evolving ecosystem of multimodal generators.

AI USE STATEMENT

LLM tools were used for language polishing and as evaluators in the experimental protocol. We also study open-weight and closed-source generative AI models as experimental targets. The authors take full responsibility for the manuscript.

ETHICS STATEMENT

This work aims to reduce unauthorized reuse of facial identity. We evaluate our method on VGGFace2 and CelebA-HQ. Human ratings were collected from external raters. They assessed facial identity similarity in generated images. Similar techniques could also be misused to evade identity-matching systems. Appendix F discusses the potential benefits, misuse risks, and mitigation measures.

REPRODUCIBILITY STATEMENT

Appendix B provides the training objective and optimization algorithm. Appendix C lists the model checkpoints, editing prompts, evaluation protocol, and implementation settings. We report random seeds and hyperparameters to support reproducibility. Our code is provided in the supplementary materials.

REFERENCES

Aniruddha Bala, Rohit Chowdhury, Rohan Jaiswal, and Siddharth Roheda. DCT-Shield: A robust frequency domain defense against malicious image editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18876–18884, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Bala_DCT-Shield_A_Robust_Frequency_Domain_Defense_against_Malicious_Image_Editing_ICCV_2025_paper.html.

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

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402, June 2023. doi: 10.1109/CVPR52729.2023.01764. URL https://openaccess.thecvf.com/content/CVPR2023/html/Brooks_InstructPix2Pix_Learning_To_Follow_Image_Editing_Instructions_CVPR_2023_paper.html.

BytePlus. Seedream 4.5: Ultra-clarity & precision in creative technology. https://www.byteplus.com/en/live/webinar20251216, December 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/hash/222dda29587fbc2979ca99fd5ed00735-Abstract-Conference.html.

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, May 2018. doi: 10.1109/FG.2018.00020. URL https://doi.org/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 *Computer Vision – ECCV 2024*, volume 15121 of *Lecture Notes in Computer Science*, pp. 126–142. Springer Nature

Switzerland, 2024. doi: 10.1007/978-3-031-73036-8_8. URL https://link.springer.com/chapter/10.1007/978-3-031-73036-8_8.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829, June 2023. doi: 10.1109/CVPR52729.2023.00276. URL https://openaccess.thecvf.com/content/CVPR2023/html/Cherti_Reproducible_Scaling_Laws_for_Contrastive_Language-Image_Learning_CVPR_2023_paper.html.

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 (ICLR)*, volume 2025, pp. 27134–27180, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/43b77cef2a83a25aa27d3271d209e4fd-Abstract-Conference.html.

Yunshu Dai, Jianwei Fei, and Fangjun Huang. IDGuard: Robust, general, identity-centric POI proactive defense against face editing abuse. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11934–11943, June 2024. doi: 10.1109/CVPR52733.2024.01134.

Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, October 2022. doi: 10.1109/TPAMI.2021.3087709. URL https://doi.org/10.1109/TPAMI.2021.3087709.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9185–9193, 2018.

Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4312–4321, 2019.

Google DeepMind. Gemini 3.1 Flash Image. https://deepmind.google/models/model-cards/gemini-3-1-flash-image/, February 2026.

Robert Hönig, Javier Rando, Nicholas Carlini, and Florian Tramèr. Adversarial perturbations cannot reliably protect artists from generative AI. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 70223–70263, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/af800ac69aea2fb5c4bcf8c6a5f3c79d-Abstract-Conference.html.

Joonsung Jeon, Woo Jae Kim, Suhyeon Ha, Sooel Son, and Sung-eui Yoon. AdvPaint: Protecting images from inpainting manipulation via adversarial attention disruption. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 76927–76940, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/bf67feee86c1711d63cb4e6dc9998702-Abstract-Conference.html.

Jaehwan Jeong, Sumin In, Sieun Kim, Hannie Shin, Jongheon Jeong, Sang Ho Yoon, Jaewook Chung, and Sangpil Kim. FaceShield: Defending facial image against deepfake threats. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10364–10374, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Jeong_FaceShield_Defending_Facial_Image_against_Deepfake_Threats_ICCV_2025_paper.html.

Xiaojun Jia, Sensen Gao, Simeng Qin, Tianyu Pang, Chao Du, Yihao Huang, Xinfeng Li, Yiming Li, Bo Li, and Yang Liu. Adversarial attacks against closed-source MLLMs via feature optimal alignment. In *Advances in Neural Information Processing Systems*, volume 38, pp. 57993–58028, 2025.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hk99zCeAb.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020.

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 (CVPR)*, pp. 15710–15718, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Le_VisiLock_Authorizing_Instruction-based_Image_editing_with_Dual_Score_Distillation_CVPR_2026_paper.html.

Chunxiao Li, Xiaoxiao Wang, Meiling Li, Boming Miao, Peng Sun, Yunjian Zhang, Xiangyang Ji, and Yao Zhu. Bridging the gap between ideal and real-world evaluation: Benchmarking ai-generated image detection in challenging scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20379–20389, 2025a.

Yuanbo Li, Tianyang Xu, Cong Hu, Tao Zhou, Xiao-Jun Wu, and Josef Kittler. Multi-paradigm collaborative adversarial attack against multi-modal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30065–30075, 2026.

Zelin Li, Ruohan Zong, Yifan Liu, Ruichen Yao, Yaokun Liu, Yang Zhang, and Dong Wang. Anti-tamper protection for unauthorized individual image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15501–15510, October 2025b. doi: 10.1109/ICCV51701.2025.01438.

Zhaoyi Li, Xiaohan Zhao, Dong-Dong Wu, Jiacheng Cui, and Zhiqiang Shen. A frustratingly simple yet highly effective attack baseline: Over 90% success rate against the strong black-box models of GPT-4.5/4o/o1. In *Advances in Neural Information Processing Systems*, volume 38, 2025c. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3ff71ee3d8ec70e172be0b2d56e73481-Abstract-Conference.html.

Hao Liang, Pietro Perona, and Guha Balakrishnan. Benchmarking algorithmic bias in face recognition: An experimental approach using synthetic faces and human evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4977–4987, October 2023. doi: 10.1109/ICCV51070.2023.00459. URL https://openaccess.thecvf.com/content/ICCV2023/html/Liang_Benchmarking_Algorithmic_Bias_in_Face_Recognition_An_Experimental_Approach_Using_ICCV_2023_paper.html.

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, 2025. URL https://arxiv.org/abs/2504.17761.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy8gdB9xx.

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.

Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. CIAGAN: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5447–5456, 2020.

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, Xunliang Cai, Yayong Guan, and Jie Hu. LongCat-Image technical report, 2025. URL https://arxiv.org/abs/2512.07584.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=aBsCjcPu_tE.

Microsoft AI Superintelligence Team. MAI-Image-2.5 launches at no. 2 for image editing on arena. https://microsoft.ai/news/introducing-mai-image-2-5/, June 2026.

Elaine M. Newton, Latanya Sweeney, and Bradley A. Malin. Preserving privacy by de-identifying face images. *IEEE Transactions on Knowledge and Data Engineering*, 17(2):232–243, 2005. doi: 10.1109/TKDE.2005.32.

OpenAI. GPT Image 2 Model. https://developers.openai.com/api/docs/models/gpt-image-2, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Tarik Can Ozden, Ozgur Kara, Oguzhan Akcin, Kerem Zaman, Shashank Srivastava, Sandeep P. Chinchali, and James Matthew Rehg. DiffVax: Optimization-free image immunization against diffusion-based editing. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=QEJaKJYOIn.

Sen Peng, Jijia Yang, Mingyue Wang, Jianfei He, and Xiaohua Jia. Protective perturbations against unauthorized data usage in diffusion-based image generation, 2024. URL https://arxiv.org/abs/2412.18791.

Qwen Team. Qwen-Image-Edit-2511: Improve consistency. https://qwen.ai/blog?id=qwen-image-edit-2511, December 2025.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model. Qwen blog, April 2026. URL https://qwen.ai/blog?id=qwen3.6-27b. Official model repository: https://huggingface.co/Qwen/Qwen3.6-27B; accessed 2026-07-25.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Thomas Roca, Anthony Cintron Roman, Jehú Torres Vega, Marcelo Duarte, Pengce Wang, Kevin White, Amit Misra, and Juan Lavista Ferres. How good are humans at detecting AI-generated images? learnings from an experiment, 2025. URL https://arxiv.org/abs/2507.18640.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1–11, 2019.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510, June 2023. doi: 10.1109/CVPR52729.2023.02155. URL https://openaccess.thecvf.com/content/CVPR2023/html/Ruiz_DreamBooth_Fine_Tuning_Text-to-Image_Diffusion_Models_for_Subject-Driven_Generation_CVPR_2023_paper.html.

Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious AI-powered image editing. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29894–29918. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/salman23a.html.

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 (CVPR)*, pp. 8631–8640, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Shao_UniDef_Universal_Defense_Against_Unauthorized_Image_Manipulation_CVPR_2026_paper.html.

Linghui Shen, Mingyue Cui, and Xingyi Yang. DeContext as defense: Safe image editing in diffusion transformers, 2025. URL https://arxiv.org/abs/2512.16625.

Yiren Song, Shengtao Lou, Xiaokang Liu, Hai Ci, Pei Yang, Jiaming Liu, and Mike Zheng Shou. Anti-Reference: Universal and immediate defense against reference-based generation, 2024. URL https://arxiv.org/abs/2412.05980.

Yiren Song, Pei Yang, Hai Ci, and Mike Zheng Shou. IDProtector: An adversarial noise encoder to protect against ID-preserving image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3019–3028, June 2025. doi: 10.1109/CVPR52734.2025.00287. URL https://openaccess.thecvf.com/content/CVPR2025/html/Song_IDProtector_An_Adversarial_Noise_Encoder_to_Protect_Against_ID-Preserving_Image_CVPR_2025_paper.html.

Sourceful. Riverflow 2.5 Fast. https://replicate.com/sourceful/riverflow-v2.5-fast, 2026.

Super Intelligence Team, Changhao Qiao, Chao Hui, Chen Li, Cunzheng Wang, Dejia Song, Jiale Zhang, Jing Li, Qiang Xiang, Runqi Wang, Shuang Sun, Wei Zhu, Xu Tang, Yao Hu, Yibo Chen, Yuhao Huang, Yuxuan Duan, Zhiyi Chen, and Ziyuan Guo. FireRed-Image-Edit-1.0 technical report, 2026. URL https://arxiv.org/abs/2602.13344.

Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N. Tran, and Anh Tran. Anti-DreamBooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2116–2127, October 2023. doi: 10.1109/ICCV51070.2023.00202. URL https://doi.org/10.1109/ICCV51070.2023.00202.

Feifei Wang, Zhentao Tan, Tianyi Wei, Yue Wu, and Qidong Huang. SimAC: A simple anti-customization method for protecting face privacy against text-to-image synthesis of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12047–12056, June 2024. doi: 10.1109/CVPR52733.2024.01145.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23806–23816, June 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_Edit_Away_and_My_Face_Will_not_Stay_Personal_Biometric_CVPR_2025_paper.html.

Zihao Wang, Yuxiang Wei, Fan Li, Renjing Pei, Hang Xu, and Wangmeng Zuo. ACE: Anti-editing concept erasure in text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23505–23515, June 2025b. doi: 10.1109/CVPR52734.2025.02189.

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, 2025. URL https://arxiv.org/abs/2508.02324.

Chenyuan Wu, Jiahao Wang, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ziyi Xia, Ze Liu, Chaofan Li, Haoge Deng, Kun Luo, Bo Zhang, Jiajun Zhang, Dong Liu, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. OmniGen2: Towards instruction-aligned multimodal generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21964–21975, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wu_OmniGen2_Towards_Instruction-Aligned_Multimodal_Generation_CVPR_2026_paper.html.

xAI. Imagine overview. https://docs.x.ai/developers/model-capabilities/imagine, 2026.

Song Xia, Meiwen Ding, Chenqi Kong, Wenhan Yang, and Xudong Jiang. Provable robustness in multimodal large language models via feature space smoothing, 2026. URL https://arxiv.org/abs/2601.16200.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2730–2739, June 2019. doi: 10.1109/CVPR.2019.00284. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Xie_Improving_Transferability_of_Adversarial_Examples_With_Input_Diversity_CVPR_2019_paper.html.

Peng Xie, Yequan Bie, Jianda Mao, Yangqiu Song, Yang Wang, Hao Chen, and Kani Chen. Chain of attack: On the robustness of vision-language models against transfer-based adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14679–14689, 2025. doi: 10.1109/CVPR52734.2025.01368.

Lizhi Xiong, Jun Li, Ziqiang Li, Weiwei Jiang, and Zhangjie Fu. No way to steal my face: Proactive defense against identity-preserving personalized generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20680–20690, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Xiong_No_Way_To_Steal_My_Face_Proactive_Defense_Against_Identity-Preserving_CVPR_2026_paper.html.

Wenkui Yang, Jie Cao, Junxian Duan, and Ran He. Towards robust defense against customization via protective perturbation resistant to diffusion-based purification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19290–19300, October 2025. doi: 10.1109/ICCV51701.2025.01793. URL https://openaccess.thecvf.com/content/ICCV2025/html/Yang_Towards_Robust_Defense_against_Customization_via_Protective_Perturbation_Resistant_to_ICCV_2025_paper.html.

Jiaming Zhang, Junhong Ye, Xingjun Ma, Yige Li, Yunfan Yang, Yunhao Chen, Jitao Sang, and Dit-Yan Yeung. Anyattack: Towards large-scale self-supervised adversarial attacks on vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19900–19909, 2025.

Xuanyu Zhang, Runyi Li, Jiwen Yu, Youmin Xu, Weiqi Li, and Jian Zhang. EditGuard: Versatile image watermarking for tamper localization and copyright protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11964–11974, 2024.

Xiaohan Zhao, Zhaoyi Li, Yaxin Luo, Jiacheng Cui, and Zhiqiang Shen. Pushing the frontier of black-box LVLM attacks via fine-grained detail targeting. *arXiv preprint arXiv:2602.17645*, 2026.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 54111–54138, 2023. doi: 10.52202/075280-2355.

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. doi: 10.1109/CVPR52733.2024.02303. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zhao_Can_Protective_Perturbation_Safeguard_Personal_Data_from_Being_Exploited_by_CVPR_2024_paper.html.

Boyang Zheng, Chumeng Liang, and Xiaoyu Wu. Targeted attack improves protection against unauthorized diffusion customization. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 17708–17738, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2d1c26fa38f31aca8712055758194bac-Abstract-Conference.html.

Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 787–796, June 2015. doi:

10.1109/CVPR.2015.7298679. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Zhu_High-Fidelity_Pose_and_2015_CVPR_paper.html.

Yao Zhu, Yuefeng Chen, Xiaodan Li, Rong Zhang, Xiang Tian, Bolun Zheng, and Yaowu Chen. Information-containing adversarial perturbation for combating facial manipulation systems. *IEEE Transactions on Information Forensics and Security*, 18:2046–2059, 2023.

A SUPPLEMENT OVERVIEW

This technical supplement provides the following supporting materials for T-VIP:

- **Methodological procedures.** Complete details of transferable identity-space construction and protected-image optimization.
- **Experimental details.** Datasets, target generators, baselines, implementation settings, prompts, and evaluation metrics.
- **Additional results.** Extended experiments on an additional dataset.
- **Extended related work.** Further discussion of relevant methods and supporting references omitted from the main paper for space.
- **Broader impact and limitation.** The potential societal benefits, limitations, misuse risks, and responsible use of T-VIP.

The main paper remains self-contained.

B METHODOLOGICAL PROCEDURES

B.1 PROJECTOR TRAINING

To amplify identity-discriminative cues while retaining each encoder's general visual representation, T-VIP augments every frozen encoder with a lightweight projection head. This subsection details how these heads are trained with identity supervision using a supervised contrastive objective (Khosla et al., 2020).

Let $\mathcal{D}_{\text{proj}} = \{(x_i, y_i)\}$ denote the identity-labeled projector-training set, where y_i is the identity of image x_i. For the m-th frozen encoder E_m, we train a linear projection head $P_m \in \mathbb{R}^{d_p \times d_m}$, where d_m is the encoder dimension and d_p is the projection dimension. We define the normalized frozen and projected features as

$$e_{m,i} = \text{norm}(E_m(x_i)), \qquad r_{m,i} = \text{norm}(P_m e_{m,i}), \tag{11}$$

where $\text{norm}(v) = v/\|v\|_2$. Only P_m is optimized, while E_m remains frozen.

Let $\mathcal{I}$ be the index set of an identity-balanced mini-batch and let $\mathcal{I}_i^+ = \{p \in \mathcal{I} \setminus \{i\} : y_p = y_i\}$ contain the positive samples for anchor i. The contrastive objective is

$$\begin{aligned} \mathcal{L}_{\text{con}}^{(m)} &= -\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \frac{1}{|\mathcal{I}_i^+|} \sum_{p \in \mathcal{I}_i^+} \log \frac{\exp\big(r_{m,i}^{\mathsf{T}} r_{m,p}/\tau_{\text{proj}}\big)}{Z_{m,i}^{\text{proj}}}, \\ Z_{m,i}^{\text{proj}} &= \sum_{a \in \mathcal{I} \setminus \{i\}} \exp\big(r_{m,i}^{\mathsf{T}} r_{m,a}/\tau_{\text{proj}}\big), \end{aligned} \tag{12}$$

where τ_{proj} is the projector-training temperature. We use a separate symbol because τ in the main paper controls the soft nearest-neighbor pull objective. Each mini-batch contains four images from each sampled identity, so every anchor has three positives and all images from other identities serve as negatives. We set $\tau_{\text{proj}} = 0.1$ in all experiments.

B.2 OVERALL PROCEDURE

The complete protected-image optimization procedure is summarized below. To reduce computational overhead, at each optimization step, z^{src} and z^{prt} are each computed from a single independently sampled crop. We implement RandomCrop using the `RandomResizedCrop` transform in torchvision with a scale range of $[0.5, 1.0]$, following prior work (Li et al., 2025c).

Algorithm 1 T-VIP protected-image optimization

Require: Source image x^{src}; identity encoder Φ; alternative identity bank $\mathcal{A}$; perturbation budget ϵ; optimization steps N

Ensure: Protected image x^{prt}

Initialize $x^{\mathrm{prt}} \leftarrow x^{\mathrm{src}}$
for $t = 1, \ldots, N$ **do**
  $z^{\mathrm{src}} \leftarrow \Phi(\mathrm{RandomCrop}(x^{\mathrm{src}}))$
  $z^{\mathrm{prt}} \leftarrow \Phi(\mathrm{RandomCrop}(x^{\mathrm{prt}}))$
  $\mathcal{L}_{\mathrm{push}} \leftarrow \cos(z^{\mathrm{prt}}, z^{\mathrm{src}})$
  $\mathcal{L}_{\mathrm{pull}} \leftarrow -\tau \log\left[\frac{1}{B}\sum_{i=1}^{B} \exp\left(\frac{\cos(z^{\mathrm{prt}}, z_i^{\mathrm{alt}})}{\tau}\right)\right]$
  $\mathcal{L}_{\mathrm{mislead}} \leftarrow \mathcal{L}_{\mathrm{push}} + \alpha\mathcal{L}_{\mathrm{pull}}$
  Update x^{prt} using $\nabla_{x^{\mathrm{prt}}}\mathcal{L}_{\mathrm{mislead}}$
  Project x^{prt} onto $\{x \in [0,1]^{H\times W\times C} : \|x - x^{\mathrm{src}}\|_\infty \le \epsilon\}$
end for
return x^{prt}

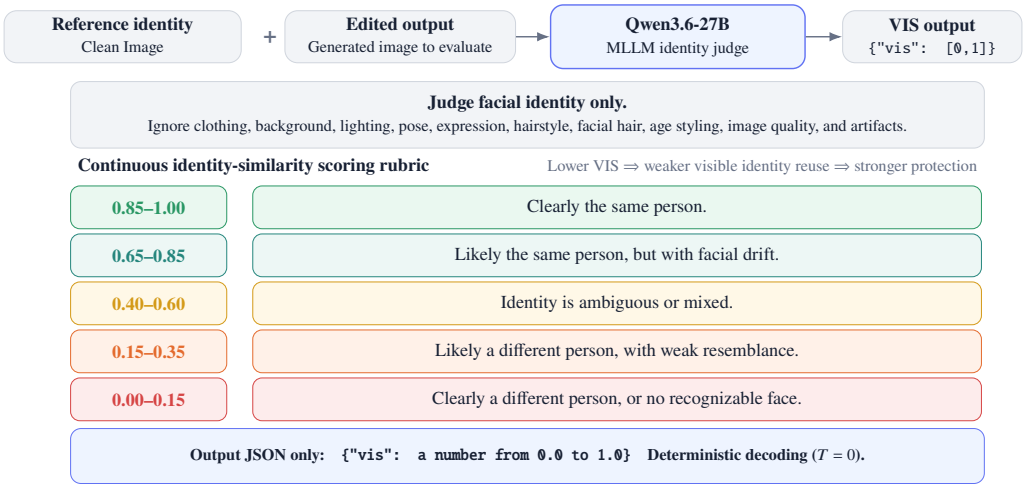

Figure 8: **VIS evaluation with a stabilized MLLM judge.** We instantiate the judge with Qwen3.6-27B (Qwen Team, 2026) and use an empirical stabilization procedure inspired by Feature-Space Smoothing (Xia et al., 2026). The wording and score intervals follow the evaluation prompt used in our implementation.

C Experimental Details

C.1 Visual Identity Similarity (VIS)

Rationale. ArcFace-based ISM inherits the sensitivities of a fixed face-recognition encoder. Changes in head pose, facial expression, or illumination can shift its embeddings even when the visible identity remains recognizable (Zhu et al., 2015; Liang et al., 2023). This ambiguity is particularly important for adversarial protection, where a low score may reflect evaluator-specific feature disruption rather than genuine suppression of identity reuse. Our proposed metric Visual Identity Similarity (VIS) instead asks an instruction-following MLLM to assess visible identity while explicitly discounting these nuisance factors. Figure 8 provides the complete judge prompt and scoring rubric. The MLLM is smoothed to be robust to perturbations (Xia et al., 2026).

Alignment with Human Judgments. We compare VIS and ISM against human judgments on 200 generated images, evenly split between open-weight and closed-source generators. Human raters assess identity similarity on a five-point scale while disregarding pose, expression, clothing, background, illumination, style, and generation artifacts. As shown in Table 6, VIS aligns more closely with human judgments than ISM across both generator groups. A small fraction of

Table 6: **Alignment with Human Judgments.** Spearman's ρ between each metric and human ratings; higher is better. Overall values use equal-size Fisher-z aggregation.

Subset	Samples	VIS	VIS w/ Smooth.	ISM
Overall	200	0.656	0.637	0.576
Open-weight	100	0.534	0.491	0.483
Closed-source	100	0.752	0.748	0.656

Table 7: **Robustness to Protective Perturbations.** Mean similarity between each source image and its protected counterpart; higher is better.

Input pair	Pairs	VIS	VIS w/ Smooth.	ISM
$x^{\mathrm{src}}, x^{\mathrm{prt}}$	50	0.027	**0.934**	0.521

open-weight generators, particularly Step1X, may carry residual protective perturbations into their outputs. Feature-space smoothing reduces the evaluator's sensitivity to these residual patterns and thereby improves the robustness of VIS.

Robustness to Perturbations. We further isolate this failure mode in Table 7 by comparing each source image directly with its protected counterpart. Raw VIS and ISM fall to 0.027 and 0.521, respectively, even though the pair depicts the same portrait. With feature-space smoothing, VIS remains 0.934. This contrast shows that smoothing substantially reduces metric exploitation by protective perturbations. In most cases, particularly for commercial generators, the perturbation itself does not survive generation. Nevertheless, this analysis highlights why identity leakage should be measured between the clean source x^{src} and the generated image x^{gen}, rather than between the protected image x^{prt} and x^{gen}. The latter allows the perturbation to directly alter the evaluator's representation of the reference and may therefore underestimate identity leakage without reflecting a genuine change in the generated identity.

Robustness to MLLM Choice. To ensure reproducibility, our primary evaluation uses a locally deployed Qwen3.6-27B judge. We additionally rescore the same 400 image pairs with GPT-5.6-sol. As shown in Table 8, the two judges exhibit strong rank agreement overall, with similarly high correlations on open-weight and closed-source generators. VIS therefore largely preserves the ordering of samples across MLLM choices, supporting its use for relative protection comparisons. GPT-5.6-sol is a more permissive judge, assigning lower VIS scores on average and more readily treating two identities as different. Qwen3.6-27B provides a more conservative test of protection in addition to being locally reproducible. T-VIP's effectiveness under this *stricter* judge further supports that its measured gains are not specific to a favorable MLLM evaluator. Taken together, our rationale, alignment, and robustness analyses motivate the broader adoption of VIS as a complementary metric for evaluating identity-protection effectiveness.

C.2 Editing Prompts

We use ten editing prompts that vary scene, clothing, gaze, or expression. Every complete prompt explicitly instructs the generator to "Preserve the person's facial identity." This provides a stringent test of identity protection, as the generator is explicitly instructed to preserve the source identity while performing the requested edit. Figure 9 shows the shared instruction and identity-preservation constraint once; concatenating them with each numbered payload gives the complete prompt.

C.3 Encoders and Target Models

Table 9 records the exact image-encoder checkpoints and target-model identifiers used in our experiments.

Table 8: **Robustness to MLLM Choice.** Agreement between Qwen3.6-27B (Q) and GPT-5.6-sol (G) VIS scores. Higher Spearman's ρ indicates stronger rank consistency; $\Delta\mu = \mu_Q - \mu_G$.

Subset	n	ρ [95% CI]	μ_Q/μ_G	$\Delta\mu$
Overall	400	0.764 [0.613, 0.882]	0.726/0.549	0.177
Open-weight	200	0.746 [0.529, 0.908]	0.684/0.549	0.135
Closed-source	200	0.777 [0.533, 0.937]	0.768/0.550	0.218

Shared instruction "Use the uploaded image as the identity reference. Generate a realistic head-and-shoulders portrait of the same person"

P01 | Sunlit park
wearing a denim jacket in a sunlit park, with a natural smile.

P02 | Beach
wearing a light linen shirt at the beach, with a relaxed neutral expression.

P03 | Neon street
wearing a black coat on a neon-lit city street at night, with a serious expression.

P04 | Office
wearing formal business attire in a modern office, with a confident expression.

P05 | Quiet cafe
wearing a dark green sweater in a quiet cafe, looking slightly thoughtful.

P06 | Snowy street
wearing a warm winter coat on a snowy street, with a gentle smile.

P07 | Art museum
wearing a beige turtleneck in an art museum, looking slightly to the left.

P08 | Library
wearing a navy cardigan in a library, looking directly at the camera.

P09 | Warm studio
wearing a plain white shirt in a warmly lit studio, laughing naturally.

P10 | Rooftop
wearing a casual blazer on a rooftop at golden hour, with a calm expression.

Shared constraint "Preserve the person's facial identity."

Figure 9: **Editing prompts used for identity-preserving generation.** Each complete prompt concatenates the shared instruction, one of the ten numbered editing payloads, and the shared identity-preservation constraint. The suite spans diverse scenes, clothing, gaze directions, and expressions while keeping the requested identity fixed.

C.4 Baselines and Fair Comparison

Table 10 summarizes why we include three baselines in the main comparison. IDProtector and ID-Guardian explicitly study cross-model or model-agnostic transfer, making them suitable black-box transfer baselines despite their original focus on personalization systems (Song et al., 2025; Xiong et al., 2026). DeContext targets in-context MM-DiT editing, the closest accessible proxy to modern closed-source editing (Shen et al., 2025). Because DeContext requires target-model gradients and separate optimization, we evaluate it only on open-weight models under white-box access.

The lower group summarizes related defenses with different reported evaluation settings (Peng et al., 2024; Van Le et al., 2023; Wang et al., 2024; Zheng et al., 2025; Yang et al., 2025; Salman et al., 2023; Chen et al., 2024; Choi et al., 2025; Jeon et al., 2025; Bala et al., 2025; Ozden et al., 2026; Wang et al., 2025a; Jeong et al., 2025). Most evaluate accessible personalization, editing, inpainting, 3D, or deepfake pipelines. UniDef generates perturbations on SD1.4 and transfers them across diffusion editing, inpainting, super-resolution, video, and 3D, without evaluating modern multimodal editors (Shao et al., 2026). Its published results therefore use different targets and protocols from ours. Anti-Reference reports transfer to SD1.5-based animation APIs (Song et al., 2024).

Model- and service-side defenses require editor modifications or provider-side verification outside our user-side threat model (Dai et al., 2024; Wang et al., 2025b; Le & Fu, 2026; Li et al., 2025b). Their access requirements and protection objectives differ from ours. We separately evaluate Anti-DreamBooth's complementarity with T-VIP.

Table 9: **Image Encoders, Evaluation Judge, and Target Models.** Exact model checkpoints and identifiers used in our experiments.

Model	Checkpoint / Model Identifier
Image encoders	
OpenAI CLIP ViT-B/16 (Radford et al., 2021)	`openai/clip-vit-base-patch16`
OpenAI CLIP ViT-B/32 (Radford et al., 2021)	`openai/clip-vit-base-patch32`
LAION CLIP ViT-B/32 (Cherti et al., 2023)	`laion/CLIP-ViT-B-32-laion2B-s34B-b79K`
DINOv2 ViT-B/14 (Oquab et al., 2024)	`facebook/dinov2-base`
Evaluation judge	
Qwen3.6-27B (Qwen Team, 2026)	`Qwen/Qwen3.6-27B`
Open-weight target models	
Qwen-Image-Edit-2511 (Qwen Team, 2025)	`ovedrive/Qwen-Image-Edit-2511-4bit`
Step1X-Edit (Liu et al., 2025)	`rkfg/Step1X-Edit-FP8`
LongCat-Image-Edit (Meituan LongCat Team et al., 2025)	`meituan-longcat/LongCat-Image-Edit`
FireRed-Image-Edit-1.1 (Super Intelligence Team et al., 2026)	`FireRedTeam/FireRed-Image-Edit-1.1`
OmniGen2 (Wu et al., 2026)	`DFloat11/OmniGen2-mllm-DF11`; `DFloat11/OmniGen2-transformer-DF11`
FLUX.1 Kontext [dev] (Black Forest Labs et al., 2025)	`black-forest-labs/FLUX.1-Kontext-dev`
Closed-source target models	
GPT Image 2 (OpenAI, 2026)	`openai/gpt-image-2`
Nano Banana 2 (Google DeepMind, 2026)	`google/gemini-3.1-flash-image`
Grok Imagine (xAI, 2026)	`x-ai/grok-imagine-image-quality`
Seedream 4.5 (BytePlus, 2025)	`bytedance-seed/seedream-4.5`
MAI-Image-2.5 (Microsoft AI Superintelligence Team, 2026)	`microsoft/mai-image-2.5`
Riverflow 2.5 Fast (Sourceful, 2026)	`sourceful/riverflow-v2.5-fast`

C.5 Implementation and Reproducibility Details

Randomness. T-VIP uses a base random seed of 2026. We initialize the Python, NumPy, and PyTorch CPU and CUDA random-number generators with this seed, and use seed $2026 + i$ for the i-th source image. This protocol makes stochastic crop sampling reproducible within a fixed hardware and software stack. For each method–input–editor combination, we generated exactly one output using the prescribed evaluation configuration. We did not perform repeated sampling, best-of-k selection, or manual cherry-picking. Commercial APIs that do not expose seed control are queried with the providers' default randomness, and their results therefore include service-side stochasticity.

Statistical Analysis. For the CelebA-HQ comparisons in Table 13, identity is the independent unit. We include every identity for which both the comparator and T-VIP outputs are available; no eligible identity-level sample is excluded. Within each method, we first average ISM and VIS separately across the ten editing prompts and six open-weight target models for each identity. We then compute the paired identity-level difference (comparator minus T-VIP) for each metric and report its mean with a 95% t confidence interval.

Computing Environment. The core T-VIP experiments were conducted on a server with eight 24-GB NVIDIA GeForce RTX 4090 GPUs and an AMD EPYC 7763 CPU.

VIS Implementation. For each evaluation instance, the judge receives a contact sheet containing up to four clean reference images of one identity, rendered with 224-pixel cells, together with the generated image. We stabilize the visual pathway using Monte Carlo feature-space smoothing. For each input image, we add $K = 8$ independently sampled Gaussian perturbations with standard deviation $\sigma = 0.25$ to the processed visual tensor. Each perturbed tensor is passed through a Qwen-specific denoiser and the vision tower, followed by a residual mapper for compatible visual features.

Table 10: **Baseline Selection and Scope.** Original protection settings and evaluation scope relative to our main comparison.

Method	Venue / Year	Original protection setting	Evaluation scope / access
Baselines included in the main comparison			
IDProtector	CVPR 2025	Encoder-based personalization	Transferability evaluated in their paper
IDGuardian	CVPR 2026	Training-free and training-based identity personalization	Explicitly studies model-agnostic protection
DeContext	arXiv 2025	In-context MM-DiT editing	Closest accessible proxy to modern closed-source editing; white-box only
Related defenses outside our main comparison			
Anti-DB, AdvDM, Mist, EUDP, SDS, MetaCloak, SimAC, ACE/ACE+, Anti-Pure	2023–2025	Diffusion personalization and customization	No evaluation on closed-source generators
PhotoGuard, EditShield, DiffusionGuard, DCT-Shield, DiffVax	2023–2026	Instruction-guided diffusion editing	No evaluation on closed-source generators
AdvPaint	ICLR 2025	Inpainting	No evaluation on closed-source generators
UniDef	CVPR 2026	Diffusion editing, inpainting, super-resolution, video, and 3D	Evaluated on SD1.x/2.x and related diffusion pipelines; modern multimodal editors not evaluated
FaceLock, FaceShield	CVPR 2025 / ICCV 2025	Portrait editing and deepfakes	No evaluation on closed-source generators
Anti-Reference	arXiv 2024	SD1.5 personalization and animation	Closed-source tests cover animation APIs, not modern multimodal image editors
IDGuard, ACE (erasure), VisiLock	2024–2026	Model-side access control and concept erasure	Requires modifying the target model
ATP	ICCV 2025	Service-assisted tamper detection	Requires provider-side verification

Table 11: **Protection Effectiveness on Open-weight Models Using CelebA-HQ.** Lower ISM and VIS indicate stronger protection. Best and second-best protection results are bolded and underlined, respectively.

Method	Qwen-Edit		Step1X		LongCat		FireRed		OmniGen2		FLUX K.	
	ISM↓	VIS↓	ISM↓	VIS↓	ISM↓	VIS↓	ISM↓	VIS↓	ISM↓	VIS↓	ISM↓	VIS↓
No protection	0.405	0.896	0.512	0.855	0.461	0.915	0.477	0.915	0.572	0.912	0.492	0.841
IDProtector	0.364	0.866	0.458	0.798	<u>0.453</u>	0.919	<u>0.418</u>	<u>0.898</u>	0.444	0.865	0.378	0.787
IDGuardian	0.374	0.879	0.487	0.813	0.495	0.914	0.432	0.917	0.511	0.888	0.387	0.778
DeContext	<u>0.360</u>	<u>0.864</u>	**0.157**	**0.015**	0.457	<u>0.858</u>	0.474	0.911	<u>0.286</u>	<u>0.335</u>	<u>0.202</u>	<u>0.518</u>
T-VIP (Ours)	**0.173**	**0.273**	<u>0.292</u>	<u>0.083</u>	**0.219**	**0.197**	**0.198**	**0.300**	**0.172**	**0.166**	**0.153**	**0.288**

We then average visual embeddings across eight draws. The resulting mean features are used in a single deterministic decoding pass with temperature zero and the fixed prompt in Figure 8. The judge returns a continuous VIS score in $[0, 1]$ as JSON; malformed or out-of-range responses are treated as invalid. We fix the smoothing seed to 12345 and cache the averaged features for identical visual tensors and smoothing configurations.

Table 12: **Protection Effectiveness on Closed-source Models Using CelebA-HQ.** Lower ISM and VIS indicate stronger protection. Best and second-best protection results are bolded and underlined, respectively.

Method	GPT Image		Nano Banana		Grok Imagine		Seedream		MAI-Image		Riverflow	
	ISM↓	VIS↓	ISM↓	VIS↓	ISM↓	VIS↓	ISM↓	VIS↓	ISM↓	VIS↓	ISM↓	VIS↓
No protection	0.689	0.938	0.623	0.812	0.670	0.938	0.710	0.967	0.651	0.945	0.498	0.951
IDProtector	<u>0.646</u>	0.925	0.602	0.807	<u>0.589</u>	<u>0.935</u>	<u>0.693</u>	<u>0.951</u>	0.645	0.939	<u>0.469</u>	0.938
IDGuardian	0.659	<u>0.922</u>	<u>0.601</u>	<u>0.799</u>	0.644	0.938	0.695	0.956	<u>0.637</u>	<u>0.932</u>	0.477	<u>0.936</u>
T-VIP (Ours)	**0.096**	**0.102**	**0.173**	**0.125**	**0.184**	**0.256**	**0.400**	**0.596**	**0.246**	**0.362**	**0.129**	**0.348**

Table 13: **Identity-level differences on open-weight CelebA-HQ evaluations.** Positive values favor T-VIP.

Comparator	ΔISM [95% CI]	ΔVIS [95% CI]
IDProtector	0.218 [0.206, 0.230]	0.638 [0.627, 0.649]
IDGuardian	0.247 [0.235, 0.258]	0.647 [0.636, 0.658]
DeContext	0.122 [0.110, 0.133]	0.366 [0.355, 0.376]

Final Hyperparameters. Each frozen encoder is paired with a linear 512-dimensional projection head trained for 500 epochs using AdamW with a learning rate of 0.001 and zero weight decay. Each mini-batch samples 256 identities and four images per identity, and the supervised contrastive temperature is $\tau_{\text{proj}} = 0.1$. For protected-image optimization, T-VIP ensembles three CLIP encoders and one DINOv2 encoder and uses an auxiliary identity bank of $B = 200$. We set the push coefficient to 1.0, the pull coefficient to $\alpha = 1.5$, and the pull temperature to $\tau = 0.04$. We optimize for 300 Adam steps with a learning rate of 0.005, apply `RandomResizedCrop` with scale range $[0.5, 1.0]$, and constrain the perturbation with $\epsilon = 16/255$ under the ℓ_∞ norm. These values are used throughout unless stated otherwise.

D Additional Results

Results on CelebA-HQ. We further evaluate cross-dataset generalization on CelebA-HQ without retraining the identity projection heads or changing the T-VIP configuration. Across the six open-weight generators, T-VIP reduces the average ISM from 0.487 without protection to 0.201 and the average VIS from 0.889 to 0.218. It also consistently outperforms the two transfer-based baselines, whose average ISM/VIS values are $0.419/0.856$ for IDProtector and $0.448/0.865$ for IDGuardian. The same pattern extends to the six closed-source generators: T-VIP lowers the average ISM from 0.640 to 0.205 and the average VIS from 0.925 to 0.298. These results show that the protection learned from the VGGFace2-based construction transfers to a distinct celebrity-face distribution across both open-weight and closed-source targets.

Distributional and Statistical Analysis. Beyond aggregate means, Table 13 summarizes the identity-level ISM and VIS differences between T-VIP and each comparator together with their 95% confidence intervals, allowing the consistency of the observed protection gains across identities to be assessed.

E Extended Related Work

E.1 Transferable Adversarial Optimization

Transferable adversarial optimization first emerged for image classifiers, where attacks sought to avoid overfitting to a single surrogate. MI-FGSM stabilizes iterative updates with gradient momentum, whereas DI^2-FGSM optimizes over randomly transformed inputs (Dong et al., 2018; Xie et al., 2019). Translation-invariant attacks aggregate gradients across shifted views, and surrogate ensembles further expose decision structure shared across architectures (Dong et al., 2018; 2019).

Recent attacks on multimodal large language models (MLLMs) adapt these principles to pretrained vision–language representations. AttackVLM aligns adversarial images with target semantics in accessible image encoders, while AnyAttack and CoA improve scalability and cross-modal consistency (Zhao et al., 2023; Zhang et al., 2025; Xie et al., 2025). M-Attack and FOA-Attack introduce local crops and global–local alignment, while MPCAttack and M-Attack-V2 diversify representations and stabilize optimization (Li et al., 2025c; Jia et al., 2025; Li et al., 2026; Zhao et al., 2026). Together, this line shifts transfer optimization from classifier decision boundaries toward semantic features shared across multimodal systems.

Relation to prior work. T-VIP builds on established methods like surrogate ensembling and input diversity. However, it extends significantly beyond general adversarial optimization. T-VIP proposes projection heads to isolate identity-discriminative cues. It also designs a novel misleading objective based on a push and pull mechanism.

E.2 General Visual Identity Protection

Beyond adversarially perturbing a portrait, visual identity can be protected at three intervention points. At release time, de-identification suppresses or replaces identity before distribution: k-Same-style aggregation provides a formal anonymity guarantee, while generative anonymization produces substitute identities that retain non-identifying information needed by downstream vision tasks (Newton et al., 2005; Maximov et al., 2020). At the model or service level, developers can encode identity-aware refusal, erase protected concepts, or gate editing behind an authorization key (Dai et al., 2024; Wang et al., 2025b; Le & Fu, 2026). After misuse, face-forgery detection and robust watermarking support attribution and tamper localization (Rössler et al., 2019; Zhang et al., 2024). These paradigms address different stages of the identity lifecycle, but impose different dependencies: release-side anonymization sacrifices portrait fidelity, model-side controls require access to the target model or serving stack, and post-hoc mechanisms document misuse without preventing generation. T-VIP targets a complementary user-side setting, preserving the visible portrait while applying protection before the downstream generator or provider is known.

F Broader Impact and Limitation

Broader impact. T-VIP may give individuals greater agency over how their portraits are reused after publication. Because protection is applied before the downstream generator is known, it can operate across platforms and closed-source services that image owners cannot inspect or control. Reducing unauthorized identity reuse could help mitigate impersonation, non-consensual editing, fraud, and reputational harm. T-VIP can also complement provider-side access control, provenance, and content moderation as part of a layered protection strategy. At the same time, similar perturbation techniques could be repurposed to evade identity-matching or provenance systems. Responsible release should therefore clearly communicate the intended defensive use and the limits of protection.

Limitation. T-VIP is designed for modern image-conditioned generators and is less effective against older training-based personalization pipelines such as DreamBooth (Ruiz et al., 2023). This boundary reflects a difference in how identity is acquired: DreamBooth fine-tunes a model on a collection of user images, whereas image-conditioned generators consume identity cues directly at inference time. Our complementarity experiment shows that this gap can be mitigated by combining T-VIP with Anti-DreamBooth (Van Le et al., 2023). The combined defense strengthens protection against DreamBooth while retaining strong protection on Qwen-Edit and Nano Banana 2. T-VIP should therefore be viewed as complementary to, rather than a replacement for, pipeline-specific defenses.