FREEID: CALIBRATING ATTENTION FOR IDENTITY PRESERVATION IN IMAGE-CONDITIONED T2I MODELS

Anonymous authors
Paper under double-blind review

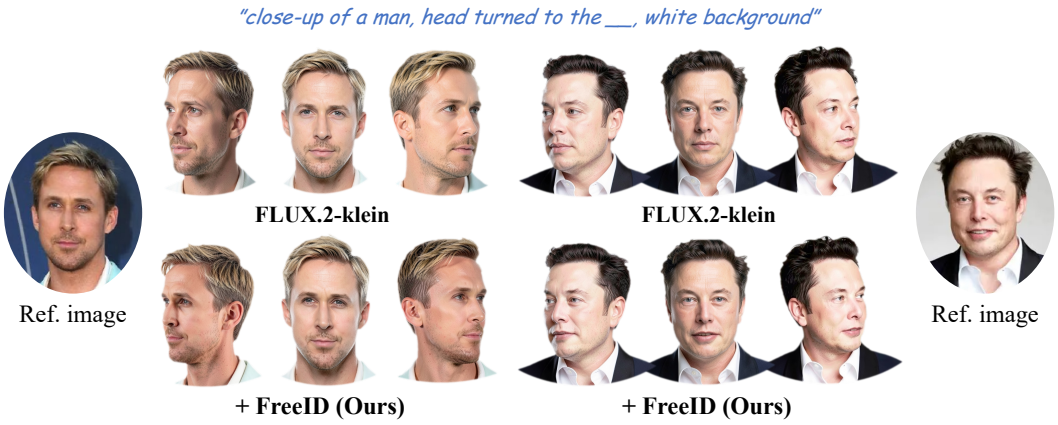


Figure 1: We propose **FreeID**, a training-free attention calibration framework that enhances identity preservation from a single reference image in image-conditioned text-to-image (IT2I) models.

ABSTRACT

Text-to-Image (T2I) diffusion models have driven rapid progress in personalized image generation, yet preserving subject identity from a reference image remains a central challenge. Conventional pipelines add identity through per-subject optimization or learned identity-conditioning modules trained on large-scale data. More recently, image-conditioned T2I (IT2I) models condition on reference images natively, processing text, reference images, and generation tokens within a unified attention framework. Yet this native conditioning still fails in demanding cases such as cross-view generation or small-scale subjects. Our analysis traces this failure to a specific mechanistic gap: IT2I models identify *which* reference tokens to attend to, but lack a mechanism to calibrate *how much* attention identity-critical features should receive, under-weighting them relative to peripheral ones; boosting facial attention alone recovers identity. Motivated by findings from human face recognition that internal facial regions carry disproportionate identity weight, we propose FreeID, a training-free framework that closes this gap at inference time. FreeID exploits the model's own attention responses to match reference facial keypoint tokens to their corresponding generation tokens, then injects sparse attention biases that amplify these identity-critical links, with magnitudes set by a closed-form bound on attention-output drift that adapts per token and step. Applied to state-of-the-art IT2I models, FreeID improves face similarity by up to 17.2% with negligible quality degradation, and further improves learned identity modules and image-to-video (I2V) models.

1 INTRODUCTION

Text-to-Image (T2I) diffusion models (Rombach et al., 2022; Esser et al., 2024; Wu et al., 2025a; Xie et al., 2025; Black Forest Labs, 2024) have demonstrated unprecedented capabilities in visual generation. This rapid advancement has sparked a growing demand for personalized image generation,

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.

which aims to synthesize images guided by specific reference images to strictly preserve subject identity within novel contexts.

In conventional T2I pipelines, reference-image conditioning is typically retrofitted onto models originally designed for text-only generation: per-subject optimization is costly, while adapter- and encoder-based alternatives rely on learned auxiliary modules, and training-free methods still operate mostly within T2I or inpainting pipelines rather than models with native reference-image tokens (Sec. 2.1). Recently, Image-conditioned Text-to-Image (IT2I) models (Black Forest Labs, 2025; Team et al., 2025; Xiao et al., 2025) have emerged as a natural alternative: they treat text, reference-image, and generation tokens as interconnected inputs within a unified attention framework, enabling direct cross-modal interactions and reference-image conditioning without additional training or auxiliary adapters.

However, even IT2I models struggle to preserve identity in challenging scenarios such as cross-view generation or when the generated subject appears at a small scale, where even subtle facial deviations can noticeably affect perceived identity. Through careful analysis, we identify what IT2I models know (Sec. 3.1) and what they do not know (Sec. 3.2). During generation, IT2I models exhibit strong semantic alignment: each generation token (query) concentrates on its corresponding reference tokens (keys) while barely attending to irrelevant regions. However, IT2I models do not know *how much* to attend to these tokens for effective identity preservation. We observe that IT2I models tend to over-attend to less identity-critical details such as hair and clothing, rather than the internal facial features that carry disproportionate identity weight—a disparity consistent with findings from human face recognition research (Bruce & Young, 1986; Sinha et al., 2006; Valentine, 1991). Building on this analysis, we propose **FreeID**, a training-free plug-in that closes this gap by calibrating the host model's reference attention toward facial keypoint tokens. Our contributions are as follows:

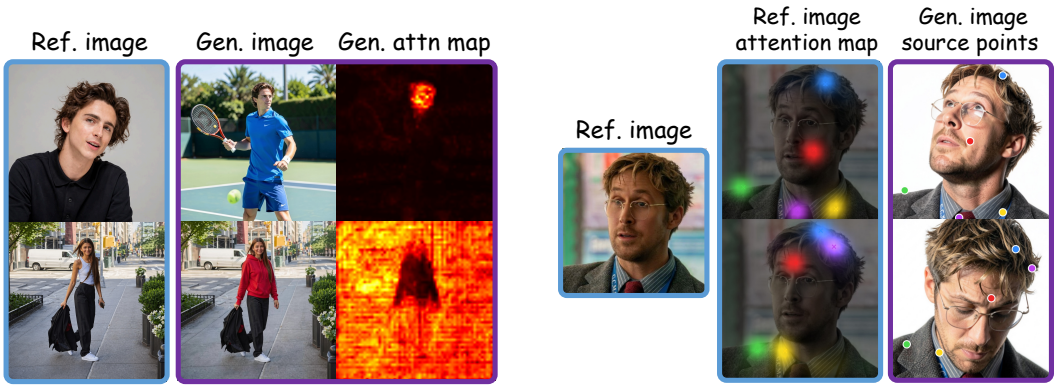
- We reveal a key limitation of IT2I models for identity preservation: their attention localizes corresponding reference regions but allocates less to identity-critical facial features than peripheral ones. Boosting facial attention alone recovers identity, confirming this under-allocation causes drift.
- We introduce FreeID, a training-free attention calibration method that leverages the model's own attention maps to match reference facial keypoint tokens to their corresponding generation tokens, then injects sparse biases only into identity-relevant reference-attention links.
- We derive a closed-form bound on the attention-output drift induced by the bias, yielding a per-token adaptive bias magnitude governed by a single parameter τ, and an architecture-aware rule that predicts τ from the number of blocks and steps as a starting point.
- We demonstrate that FreeID improves identity preservation across diverse IT2I backbones with minimal impact on image quality and text alignment, and complements learned identity modules and image-to-video models.

2 Related Work

2.1 Identity preservation in diffusion models

Early personalization methods learn specific concepts from reference images through either model weight fine-tuning (Gu et al., 2023; Kumari et al., 2023; Liu et al., 2023; Ruiz et al., 2023; Shah et al., 2024; Zhu et al., 2025; Seo et al., 2025), which updates the network weights directly, or embedding optimization (Gal et al., 2023), which learns a new textual representation without altering the base model. While these techniques achieve identity preservation, they require significant computational resources and time due to the iterative optimization process necessary for each new subject.

To improve efficiency, encoder-based methods have been developed to pre-train auxiliary modules on large-scale datasets (Li et al., 2023; Zhang et al., 2024a; Wang et al., 2025; Li et al., 2024; Guo et al., 2024; Zhang et al., 2024b). Instead of training full encoders, some approaches learn specialized adapters to inject identity (Chen et al., 2024; Wang et al., 2024; Tan et al., 2025; Ye et al., 2023). While these approaches enable efficient zero-shot inference, they require large-scale identity-specific training, and FreeID can be applied on top of them (Sec. 5.4). An alternative line of work sidesteps external modules by fine-tuning the base model itself on self-generated paired data (Cai et al., 2025),



(a) Region awareness (b) Semantic alignment

Figure 2: **Intrinsic spatial awareness in IT2I models. (a)** Attention maps showing which regions of the generated image are most influenced by the reference image. **(b)** Localized semantic alignment analysis revealing that IT2I models identify which regions of the reference image to attend to for each local area in the generated output.

achieving competitive zero-shot identity preservation without per-subject optimization—but still requiring a training stage.

Training-free methods offer an efficient alternative via shared attention and feature injection (Tewel et al., 2024), inpainting-based prompting with uniform reference-attention enhancement (Shin et al., 2025), token replacement (Feng et al., 2026), sampling-time concept composition (Kwon & Ye, 2025), or attention manipulation for pretrained ID adapters (Karpukhin et al., 2025). While avoiding per-subject parameter updates, they generally retrofit reference conditioning onto T2I/inpainting pipelines or adapters, rather than calibrating identity-relevant reference attention within native IT2I models.

2.2 Image-conditioned text-to-image models

Recently, Image-conditioned Text-to-Image (IT2I) models (Black Forest Labs, 2025; Team et al., 2025; Xiao et al., 2025) have emerged based on unified diffusion transformers. These models natively process multi-modal inputs. As this paradigm is nascent, the number of publicly available IT2I models remains limited; several capable systems exist only as closed APIs (OpenAI, 2026; Google, 2026; xAI, 2026; ByteDance Seed, 2026), where the underlying architectures are not disclosed.

IT2I models support visual conditions through a unified token sequence. Concatenating text tokens $\mathbf{T}$, reference image tokens $\mathbf{I}$, and generation tokens $\mathbf{Z}$, the update for generation-token queries is:

$$\mathbf{Z} \leftarrow \mathrm{softmax}_{\mathrm{key}}\left(\frac{(\mathbf{Z}\mathbf{W}_Q)([\mathbf{T};\mathbf{I};\mathbf{Z}]\mathbf{W}_K)^\top}{\sqrt{d_h}}\right)([\mathbf{T};\mathbf{I};\mathbf{Z}]\mathbf{W}_V). \tag{1}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are learned projection matrices and d_h is the head dimension. Queries from $\mathbf{Z}$ thus interact directly with the keys and values of both $\mathbf{T}$ and $\mathbf{I}$, whereas standard T2I models use only $[\mathbf{T};\mathbf{Z}]$ and require additional encoders, adapters, or fine-tuning to incorporate reference images.

3 Analysis

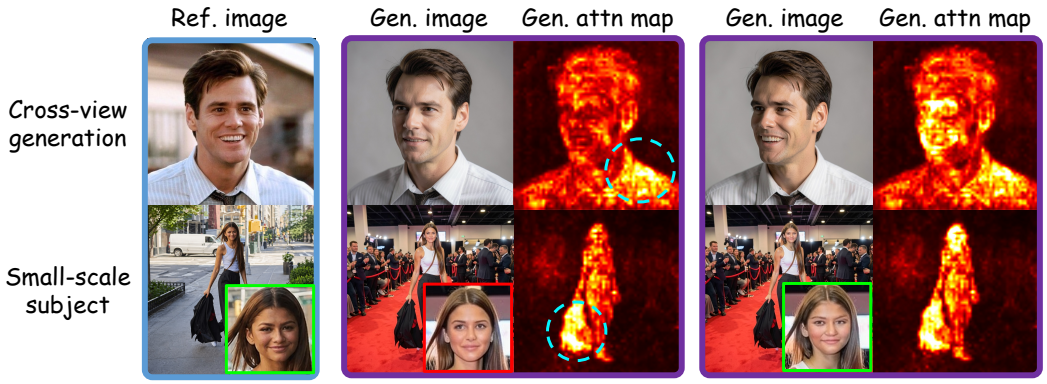
3.1 What IT2I models already know: Region awareness and semantic alignment

Fig. 2(a) presents attention maps that indicate which regions of the generated image are most influenced by the reference image. Each attention value represents the aggregated influence across all reference latent tokens, demonstrating that IT2I models understand *which* spatial regions should be derived from the reference image. This observation is supported by quantitative analysis over 50 samples. We measure each generation token's reference attention, normalized to its sample's mean. The result shows higher reference attention in personalized regions (2.56) than in background regions

(0.66), indicating that personalized regions rely more strongly on the reference image. Full protocol and metric definitions are in Appendix E.2.

While Fig. 2(a) provides an aggregated view of how strongly each generated region attends to the reference image, we analyze *where* in the reference image each local region of the generated output attends to. We first quantify the selectivity of reference attention using normalized attention entropy, finding lower entropy in personalized regions (0.52) than in background regions (0.70), indicating that personalized regions concentrate their attention on a smaller set of reference tokens. Fig. 2(b) reveals the specific reference locations that each generated region relies on. IT2I models exhibit semantic alignment: each local region attends predominantly to its corresponding region in the reference image. This attention is fine-grained, even within visually similar or repetitive regions such as hair.

3.2 What IT2I models do not know: Misaligned attention allocation for identity preservation



(a) Flux.2-klein (b) **FreeID (ours)**

Figure 3: **Misaligned attention allocation in IT2I models. (a)** Attention maps revealing that IT2I models disproportionately focus on peripheral features (hair, clothing) rather than identity-defining facial features (highlighted in cyan). **(b)** Effect of attention bias injection: applying attention bias to facial regions during intermediate denoising steps improves identity preservation but requires careful calibration to avoid artifacts.

Fig. 3(a) shows attention maps during identity-preserved image generation, similar to Fig. 2(a), but reveals an imbalance in attention allocation: IT2I models know *where* to attend, but not *how much*. Cognitive research in face recognition (Bruce & Young, 1986; Sinha et al., 2006) shows that humans perceive identity primarily through internal facial features rather than peripheral details such as hair texture or clothing, suggesting that these features should receive the most reference attention. However, the attention maps reveal that current IT2I models disproportionately attend to these peripheral elements. Our quantitative analysis supports this: decomposing the personalized region into face and peripheral, we observe higher per-token reference attention in peripheral regions (2.73) than in face regions (1.78). This imbalance explains why identity drift concentrates in the two settings of Fig. 3(a): peripheral regions such as hair and clothing can be copied from the reference regardless of pose, whereas under viewpoint change the face must be re-synthesized from weakly attended tokens, and at small scale those tokens are few to begin with.

Fig. 3(b) demonstrates the effect of injecting attention bias toward facial regions during intermediate denoising steps, using the calibration we develop in Sec. 4. When attention bias is applied to guide the model toward face-relevant reference tokens, identity preservation improves compared to the baseline. However, this injection requires calibration: excessive bias values or attending to undesired reference regions can cause image degradation, introducing artifacts or incorporating background elements into the generated subject.

These observations reveal a critical gap in current IT2I models: while they possess strong spatial correspondence capabilities, they lack the ability to allocate sufficient attention to identity-critical facial features. Importantly, this under-allocation is causal rather than merely correlated with identity drift: Fig. 3(b) shows that raising facial attention alone, with the model itself unchanged, recovers

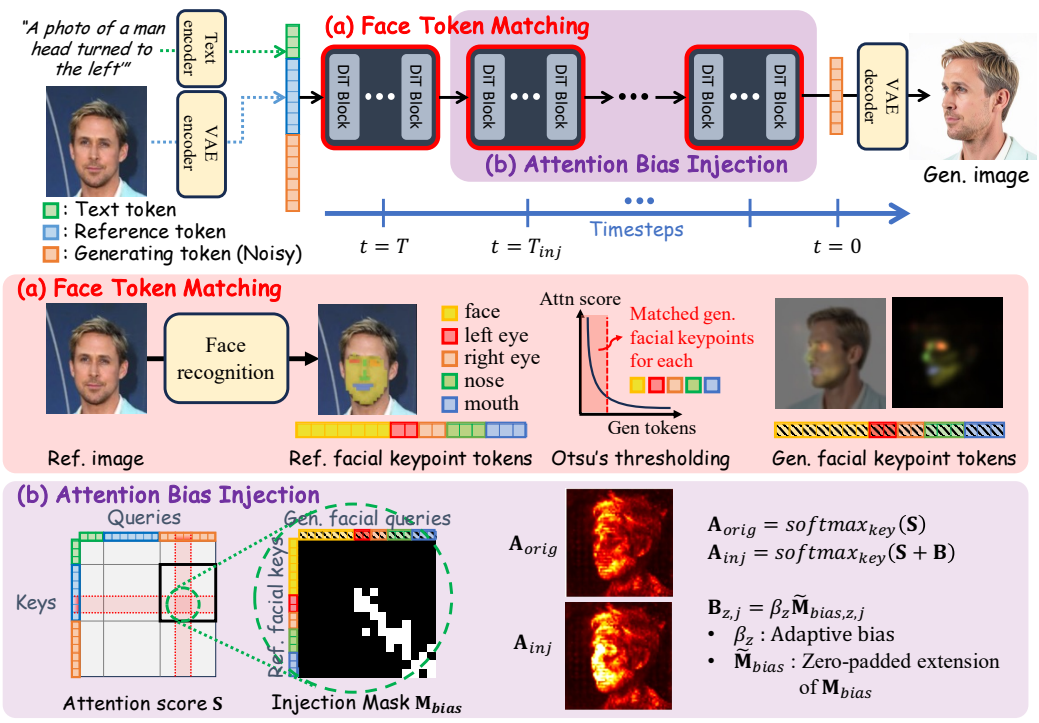


Figure 4: **Overall framework of FreeID.** Text, reference image, and generating tokens are concatenated as multimodal inputs to the IT2I model. FreeID enhances identity preservation performance through two key stages. **(a)** We detect facial keypoints in the reference image and match them to corresponding generating tokens through attention analysis, without requiring VAE decoding. **(b)** We identify the reference tokens most attended to by each facial keypoint generation token, construct an injection mask, and apply adaptive bias to amplify attention weights. This process increases attention scores for identity-critical facial features (see Sec. 3.2), resulting in improved identity preservation.

identity. In the following section, we introduce FreeID, a training-free framework that closes this gap by directly calibrating reference attention toward identity-critical facial regions.

4 Method

Based on Sec. 3.1 and 3.2, we propose **FreeID**, a training-free method to enhance identity preservation in IT2I models. FreeID consists of two components: face token matching (Sec. 4.1) for identifying face tokens, and attention bias injection (Sec. 4.2) for strengthening identity consistency.

4.1 Face token matching

As observed in Sec. 3.1, generation tokens in IT2I models already attend to corresponding regions of the reference image. We exploit this property to identify key facial identity tokens.

Let $\mathbf{S} \in \mathbb{R}^{N_Z \times N}$ be the pre-softmax logits from generation-token queries to the full key sequence $[T, I, Z]$. We compute $\mathbf{A} = \mathrm{softmax}_{\mathrm{key}}(\mathbf{S})$ and denote its reference-token slice by $\mathbf{A}^I = \mathbf{A}_{:,\mathcal{I}} \in \mathbb{R}^{N_Z \times N_I}$. A face recognition model detects K semantic facial keypoints in the reference image, and each keypoint category k is mapped to reference token indices $\mathcal{R}_k$. For each keypoint, we aggregate the attention received by its reference tokens:

$$\mathbf{M}_k = \sum_{i \in \mathcal{R}_k} \mathbf{A}^I_{:,i}. \tag{2}$$

The response map $\mathbf{M}_k$ is reshaped to the generation-token grid and thresholded by Otsu's method (Otsu, 1979) to obtain keypoint-aligned generation tokens at each denoising step (ablated against fixed thresholds in Appendix D.6).

4.2 ATTENTION BIAS INJECTION

Although generation tokens attend to the corresponding reference regions, their attention strength is often insufficient for identity preservation. We therefore amplify only the most relevant face-related links, since globally amplifying reference tokens can introduce artifacts such as color distortion (Fig. 6).

For each keypoint-aligned generation token, we sort its reference attention weights and retain only the top-ranked reference tokens whose cumulative mass reaches $r = 0.8$ of its total reference attention. This yields a binary mask $\mathbf{M}_{\text{bias}} \in \{0,1\}^{N_Z \times N_I}$. We then inject an adaptive bias matrix $\mathbf{B}$ with nonzero entries restricted by its zero-padded extension $\widetilde{\mathbf{M}}_{\text{bias}} \in \{0,1\}^{N_Z \times N}$:

$$\mathbf{A}_{\text{inj}} = \text{softmax}_{\text{key}}(\mathbf{S} + \mathbf{B}). \tag{3}$$

Here, $\mathbf{B}_{z,j} = \beta_z \widetilde{\mathbf{M}}_{\text{bias},z,j}$, where the token-wise intensity β_z is adaptively selected to avoid over-amplification. Attention manipulation has been used for text-driven editing (Hertz et al., 2023) and layout control via additive logit modulation (Kim et al., 2023), but with user-specified targets on text-token attention; our bias instead acts on reference attention at model-found correspondences. We detail the adaptive bias regulation in Sec. 4.3.

4.3 ADAPTIVE BIAS SELECTION

To prevent image degradation from excessive bias injection, we regulate β_z by bounding the attention-output drift caused by the injected bias. For a generation token z, let Ω_z be the selected key positions activated by $\widetilde{\mathbf{M}}_{\text{bias}}$, and let $p_\Omega^{(0)}(z)$ be their baseline attention probability. When the selected logits are increased by β_z and all other logits are unchanged, the selected probability mass becomes:

$$p_\Omega^{(\beta_z)}(z) = \frac{e^{\beta_z} p_\Omega^{(0)}(z)}{e^{\beta_z} p_\Omega^{(0)}(z) + (1 - p_\Omega^{(0)}(z))}. \tag{4}$$

Let $\Delta p_z = p_\Omega^{(\beta_z)}(z) - p_\Omega^{(0)}(z)$. Under the two-group view of selected keys versus their complement, the resulting attention-output perturbation is bounded by:

$$\|\Delta \mathbf{o}_z\| = \Delta p_z \cdot \|\mathbf{v}_\Omega - \mathbf{v}_{\bar{\Omega}}\| \leq \tau, \tag{5}$$

where $||\cdot||$ denotes the unscaled L_2 norm over the attention head dimension (d_h), and $\mathbf{v}_\Omega$ and $\mathbf{v}_{\bar{\Omega}}$ are the expected context values of the selected keys and their complement, respectively. We obtain the maximum β_z by solving Eq. 5 independently for each localized region and denoising step. This adaptive rule improves identity preservation while limiting image degradation. Detailed derivation of Eq. 5 and maximum β_z is in the Appendix A.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We evaluate FreeID on three IT2I backbones: FLUX.2-dev (Black Forest Labs, 2025), FLUX.2-klein (Black Forest Labs, 2025), and LongCat-Image-Edit-Turbo (Team et al., 2025). We conducted our experiments using the CelebA-HQ (Karras et al., 2018) dataset and a self-curated dataset sourced from publicly available portrait images on Unsplash (Unsplash, 2026). We measure InsightFace (InsightFace AI Technology, 2021) and AdaFace (Kim et al., 2022) for face similarity, CLIP-IQA (Wang et al., 2023) for image quality, and CLIP-T (Radford et al., 2021) for text alignment. Additional details are provided in the Appendix B. FreeID adds 15.5% latency on average with no additional memory (Appendix F).

5.2 MAIN RESULTS

Qualitative results. Fig. 5 compares the IT2I backbones with their FreeID-enhanced counterparts. Generated faces without FreeID can drift from the reference identity, particularly when the viewpoint differs from the reference image or the face is at a small scale. FreeID better preserves facial

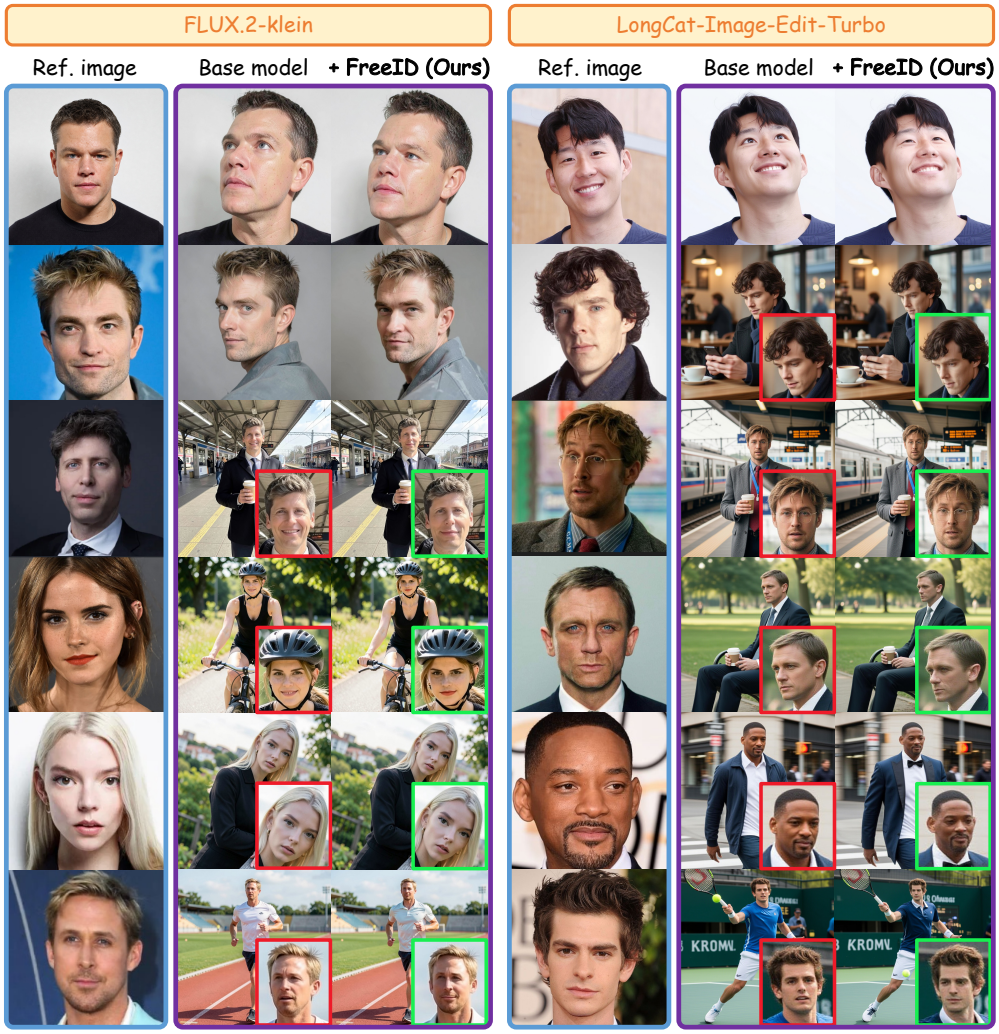


Figure 5: **Qualitative comparison on IT2I backbones.** FreeID improves identity preservation over the original IT2I models in challenging settings, including viewpoint changes and small face scales.

characteristics such as face shape, eye region, nose structure, and other local facial details under limited facial resolution or substantial view transformation. Importantly, the generated images still preserve the backbone's ability to model lighting, background, and scene layout, while following prompt-specified changes in expression, viewpoint, and context.

Quantitative results. Tab. 1 reports face similarity, image quality, and text-image alignment on Unsplash and CelebA across three IT2I backbones. FreeID consistently improves InsightFace and AdaFace over the corresponding host models on both datasets, showing that inference-time attention calibration strengthens identity preservation without additional training or model-specific fine-tuning. Meanwhile, CLIP-T remains nearly unchanged, and CLIP-IQA decreases modestly or even increases, suggesting that text alignment and overall generation quality are largely preserved.

Controlled comparison on the same backbone. On the same FLUX.2 backbones, FreeID outperforms per-subject fine-tuning (DreamBooth) and training-free token replacement (Personalize Anything): on FLUX.2-dev, InsightFace reaches 0.749 versus 0.588 and 0.322 (Table A5), comparable to DreamO (0.758, FLUX.1) without identity-specific training. On an expanded set of 200 Unsplash identities, the improvement persists (+15.8%, 95% CI strictly positive) and holds across viewpoint, face scale, reference resolution, and demographic subgroups (Appendix D.2).

Table 1: **Quantitative comparison.** We evaluate the effectiveness of FreeID across three different IT2I models. "+ FreeID" denotes the application of our method to each host model. We report the 95% confidence intervals (CI) for the improvements over each host model in Appendix B.6, and comparisons with personalization methods under the same backbones in Table A5.

Model	Face similarity		Image quality	Text alignment
	InsightFace (↑)	AdaFace (↑)	CLIP-IQA (↑)	CLIP-T (↑)
Dataset: Unsplash				
FLUX.2-klein	0.478	0.446	0.655	0.236
+ **FreeID**	0.537 (+12.3%)	0.504 (+13.0%)	0.656 (+0.2%)	0.236 (0.0%)
FLUX.2-dev	0.709	0.670	0.652	0.234
+ **FreeID**	0.749 (+5.6%)	0.710 (+6.0%)	0.651 (-0.2%)	0.234 (0.0%)
LongCat-Image-Edit-Turbo	0.633	0.589	0.664	0.236
+ **FreeID**	0.665 (+5.1%)	0.624 (+5.9%)	0.664 (0.0%)	0.236 (0.0%)
Dataset: CelebA				
FLUX.2-klein	0.417	0.401	0.670	0.231
+ **FreeID**	0.485 (+16.3%)	0.470 (+17.2%)	0.670 (0.0%)	0.230 (-0.4%)
FLUX.2-dev	0.639	0.621	0.676	0.230
+ **FreeID**	0.678 (+6.1%)	0.663 (+6.8%)	0.677 (+0.1%)	0.229 (-0.4%)
LongCat-Image-Edit-Turbo	0.511	0.494	0.700	0.233
+ **FreeID**	0.546 (+6.8%)	0.529 (+7.1%)	0.698 (-0.3%)	0.232 (-0.4%)

User study. We conduct a user study with 114 participants to evaluate perceived identity preservation across the original IT2I backbones. Participants are shown paired results from an original backbone and its FreeID-enhanced counterpart, selecting the image that better preserves the identity of the reference person. As shown in Tab. 2, FreeID receives substantially higher preference than the corresponding baseline across all evaluated backbones. This complements the quantitative results, as perceived identity depends on subtle facial cues that face-similarity metrics may fail to fully capture.

Table 2: Preference rates for facial identity preservation.

Model	**Pref. (%)**
FLUX.2-klein	9.94
+ **FreeID**	90.06
FLUX.2-dev	18.75
+ **FreeID**	81.25
LongCat	12.74
+ **FreeID**	87.26

5.3 Ablation Studies

Which generation tokens should be biased? As shown in Fig. 6 (top), applying the attention bias to all generation tokens degrades identity preservation. This setting allows identity-oriented reference guidance to affect non-face regions, which reduces the spatial locality of the intervention and can distort local appearance. Unlike region-agnostic reference-attention enhancement (Shin et al., 2025), restricting the bias to face-aligned generation tokens concentrates calibration where identity is needed, leaving non-face regions to the generative prior.

How many reference tokens should be amplified? We selectively amplify top-ranked reference tokens up to a cumulative attention mass r. Fig. 6 (bottom) compares our cumulative thresholding strategy with full reference-token amplification. When all reference tokens are amplified ($r = 1.0$), the selected generation tokens increase their dependence not only on identity-critical facial tokens but also on low-relevance long-tail reference tokens, such as background, hair, clothing, and color-related regions. This can introduce artifacts such as face color contamination. By setting the threshold, we discard these noisy tokens and concentrate attention bias solely on identity-critical features, preserving both identity fidelity and image quality.

Effect of adaptive bias value. Fig. 7 shows the trade-off between face similarity and overall image quality according to the upper bound τ. While increasing τ initially boosts identity preservation

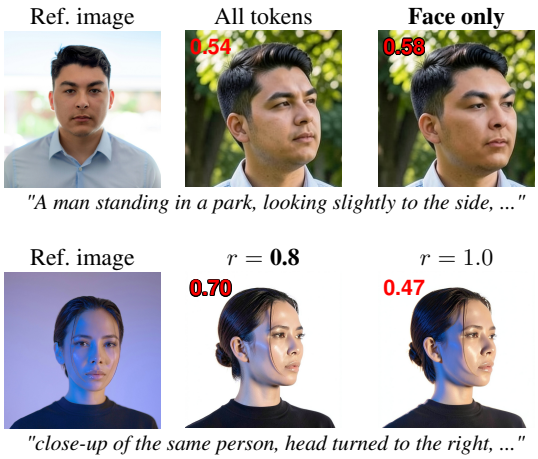


Figure 6: **Case studies for attention tokens. (top)** Applying bias to all gen. tokens degrades identity resemblance, whereas targeting only face tokens resolves this. Generated images are cropped for visualization. **(bottom)** Amplifying all ref. tokens ($r = 1.0$) causes color contamination, which is mitigated by thresholding ($r = 0.8$). Red numbers represent the InsightFace values.

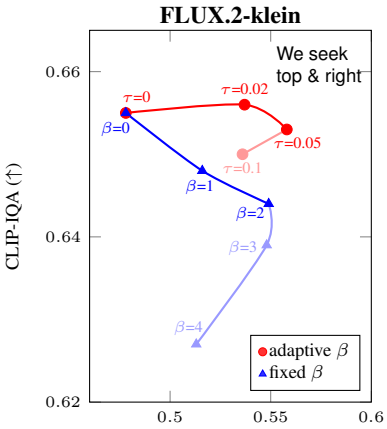


Figure 7: **Effect of adaptive bias**. By limiting attention output perturbation τ, adaptive bias (red) achieves a better Pareto front—improving face similarity with less quality drop compared to fixed bias β (blue). However, excessively large bias degrades face similarity and image quality in both settings.

with negligible cost to image quality, an excessively high τ over-perturbs the latent space, ultimately degrading both metrics. A first-order approximation of the cumulative perturbation across blocks and steps motivates $\tau \approx 1.5/(N_{\text{blocks}} N_{\text{steps}})$, which closely follows the empirically selected τ for all three backbones (Appendix B.3, Table A1) and serves as a starting point for new architectures.

5.4 Generalization

Compatibility with existing T2I-based personalization methods. We evaluate FreeID on PuLID (Guo et al., 2024) and DreamO (Mou et al., 2025), strong image personalization/customization methods with dedicated identity-conditioning mechanisms. As shown in Tab. 3, FreeID improves both, reaching 0.804 with DreamO, showing that it complements rather than competes with learned identity modules.

Extension to image-to-video generation. We also apply FreeID to image-to-video (I2V) models (Wan et al., 2025; Kong et al., 2024), where identity preservation is evaluated across generated frames from a reference image and text instruction. FreeID calibrates reference-attention links during inference while keeping the video model weights fixed.

Table 3: **Generalization across diverse generation tasks.** We evaluate the versatility of FreeID by applying it as a purely inference-time strategy to both T2I-based personalization/customization frameworks and I2V models.

	Model	InsightFace (↑)	AdaFace (↑)
T2I	PuLID	0.721	0.677
	+ **FreeID**	**0.735**	**0.685**
	DreamO	0.758	0.736
	+ **FreeID**	**0.804**	**0.786**
I2V	Wan2.2	0.752	0.723
	+ **FreeID**	**0.780**	**0.755**
	Hunyuan-I2V	0.786	0.768
	+ **FreeID**	**0.839**	**0.826**

6 Conclusion

We introduced FreeID, a training-free framework for improving identity preservation in image-conditioned text-to-image models. We showed that generation tokens exhibit fine-grained semantic alignment with reference regions, but their attention is not sufficiently concentrated on identity-critical facial features, especially under challenging viewpoint changes or small-face generation. FreeID closes this gap by matching facial keypoints to face-aligned generation tokens and selectively strengthening their connections to identity-relevant reference tokens, with the bias magnitude bounded by the induced attention-output drift. Experiments across multiple IT2I backbones show consistent gains in identity similarity, which carry over to learned personalization and I2V pipelines.

AI USE STATEMENT

In this work, we used generative AI tools to assist with parts of mathematical proof writing, method implementation, and translation. We have not used generative AI tools for other applicable tasks requiring disclosure, and the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to summarize and analyze existing literature. We have reviewed all AI-assisted work: mathematical derivations were manually verified by the authors, AI-assisted code was tested for correctness, translations were reviewed against the original text, and literature summaries were cross-checked with the original sources. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our work focuses on identity-preserving image generation, which may raise concerns regarding privacy and potential misuse of generated content. We use publicly available datasets and images solely for research and evaluation purposes. Our user study was conducted with informed participation and collected only responses necessary for evaluating perceived identity preservation. We encourage responsible use of the proposed method and do not intend it for deceptive or harmful applications.

REPRODUCIBILITY STATEMENT

We provide detailed derivations of our method in Appendix A and implementation details, model configurations, and hyperparameters in Appendix B. The evaluation protocol, metrics, and prompts used in our experiments are also documented in the appendix. These details are provided to facilitate reproduction of the reported results.

REFERENCES

Black Forest Labs. FLUX.1. https://github.com/black-forest-labs/flux, 2024.

Black Forest Labs. FLUX.2. https://github.com/black-forest-labs/flux2, 2025.

Vicki Bruce and Andy Young. Understanding face recognition. *British journal of psychology*, 77(3): 305–327, 1986.

ByteDance Seed. Seedream. https://seed.bytedance.com/en/seedream4_5, 2026.

Shengqu Cai, Eric Ryan Chan, Yunzhi Zhang, Leonidas Guibas, Jiajun Wu, and Gordon Wetzstein. Diffusion self-distillation for zero-shot customized image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18434–18443, 2025.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025.

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *CVPR*, 2024.

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 *ICML*, 2024.

Haoran Feng, Zehuan Huang, Lin Li, and Lu Sheng. Personalize anything for free with diffusion transformer. In *AAAI*, 2026.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023.

Google. Nano Banana. https://deepmind.google/models/gemini/, 2026.

Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In *NeurIPS*, 2023.

Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, Peng Zhang, and Qian He. Pulid: Pure and lightning id customization via contrastive alignment. In *NeurIPS*, 2024.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *ICLR*, 2023.

InsightFace AI Technology. InsightFace. `https://github.com/deepinsight/insightface`, 2021.

Sergey Karpukhin, Vadim Titov, Andrey Kuznetsov, and Aibek Alanov. Fastface: Tuning identity preservation in distilled diffusion via guidance and attention. *arXiv preprint arXiv:2505.21144*, 2025.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018.

Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *CVPR*, 2022.

Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023.

Gihyun Kwon and Jong Chul Ye. Tweediemix: Improving multi-concept fusion for diffusion-based image/video generation. In *ICLR*, 2025.

Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *NeurIPS*, 2023.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *CVPR*, 2024.

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*, 2025.

Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: concept neurons in diffusion models for customized generation. In *ICML*, 2023.

Chong Mou, Yanze Wu, Wenxu Wu, Zinan Guo, Pengze Zhang, Yufeng Cheng, Yiming Luo, Fei Ding, Shiwen Zhang, Xinghui Li, et al. Dreamo: A unified framework for image customization. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025.

OpenAI. GPT Image. `https://chatgpt.com/images/`, 2026.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 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 *CVPR*, 2023.

Hoigi Seo, Wongi Jeong, Kyungryeol Lee, and Se Young Chun. Efficient personalization of quantized diffusion model without backpropagation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7717–7727, 2025.

Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. In *ECCV*, 2024.

Chaehun Shin, Jooyoung Choi, Heeseung Kim, and Sungroh Yoon. Large-scale text-to-image model with inpainting is a zero-shot subject-driven image generator. In *CVPR*, 2025.

Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94 (11):1948–1962, 2006.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *ICCV*, 2025.

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, et al. Longcat-image technical report. *arXiv preprint arXiv:2512.07584*, 2025.

Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. In *ACM TOG*, 2024.

Unsplash. Unsplash dataset. `https://unsplash.com/`, 2026. Accessed: 2026-04-23.

Tim Valentine. A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology Section A*, 43(2):161–204, 1991.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models, 2025. URL `https://arxiv.org/abs/2503.20314`.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.

Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. In *ICLR*, 2025.

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*, 2025a.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025b.

xAI. grok-imagine-image. https://grok.com/imagine, 2026.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *CVPR*, 2025.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *ICLR*, 2025.

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.

Z.ai. GLM-Image. https://z.ai/blog/glm-image, 2026.

Shilong Zhang, Lianghua Huang, Xi Chen, Yifei Zhang, Zhi-Fan Wu, Yutong Feng, Wei Wang, Yujun Shen, Yu Liu, and Ping Luo. Flashface: Human image personalization with high-fidelity identity preservation. *arXiv preprint arXiv:2403.17008*, 2024a.

Yuxuan Zhang, Yiren Song, Jiaming Liu, Rui Wang, Jinpeng Yu, Hao Tang, Huaxia Li, Xu Tang, Yao Hu, Han Pan, et al. Ssr-encoder: Encoding selective subject representation for subject-driven generation. In *CVPR*, 2024b.

Chenyang Zhu, Kai Li, Yue Ma, Chunming He, and Xiu Li. Multibooth: Towards generating all your concepts in an image from text. In *AAAI*, 2025.

A PROOFS

Derivation of Eq. 5. We derive Eq. 5 for a fixed denoising step, layer, head, and generation token z; these indices are omitted for clarity. Let s_j be the pre-softmax attention logit from query z to key j, and let $\mathbf{v}_j$ be the corresponding value vector. Let Ω_z denote the selected key positions activated by $\widetilde{\mathbf{M}}_{\text{bias}}$, and let $\bar{\Omega}_z$ be its complement.

The baseline attention weight is

$$a_j^{(0)} = \frac{\exp(s_j)}{\sum_k \exp(s_k)}. \tag{A1}$$

We define the baseline probability mass assigned to the selected positions as

$$p_\Omega^{(0)}(z) = \sum_{j \in \Omega_z} a_j^{(0)}. \tag{A2}$$

When a non-negative bias β_z is added uniformly to all selected logits, the new attention weight becomes

$$a_j^{(\beta_z)} = \frac{\exp(s_j + \beta_z \mathbf{1}[j \in \Omega_z])}{\sum_k \exp(s_k + \beta_z \mathbf{1}[k \in \Omega_z])}. \tag{A3}$$

Therefore, the updated probability mass on Ω_z is

$$p_\Omega^{(\beta_z)}(z) = \sum_{j \in \Omega_z} a_j^{(\beta_z)} = \frac{e^{\beta_z} p_\Omega^{(0)}(z)}{e^{\beta_z} p_\Omega^{(0)}(z) + (1 - p_\Omega^{(0)}(z))}. \tag{A4}$$

Next, define the expected context values of the selected keys and their complement as

$$\mathbf{v}_\Omega = \frac{1}{p_\Omega^{(0)}(z)} \sum_{j \in \Omega_z} a_j^{(0)} \mathbf{v}_j, \qquad \mathbf{v}_{\bar{\Omega}} = \frac{1}{1 - p_\Omega^{(0)}(z)} \sum_{j \notin \Omega_z} a_j^{(0)} \mathbf{v}_j. \tag{A5}$$

Since the same bias β_z is added to every logit in Ω_z, the relative attention distribution within Ω_z is unchanged. The same holds for its complement. Thus, the attention outputs before and after bias injection can be written as

$$\mathbf{o}_z^{(0)} = p_\Omega^{(0)}(z) \mathbf{v}_\Omega + (1 - p_\Omega^{(0)}(z)) \mathbf{v}_{\bar{\Omega}}, \tag{A6}$$

and

$$\mathbf{o}_z^{(\beta_z)} = p_\Omega^{(\beta_z)}(z) \mathbf{v}_\Omega + (1 - p_\Omega^{(\beta_z)}(z)) \mathbf{v}_{\bar{\Omega}}. \tag{A7}$$

Subtracting the two outputs gives

$$\Delta \mathbf{o}_z = \mathbf{o}_z^{(\beta_z)} - \mathbf{o}_z^{(0)} = \Delta p_z \left(\mathbf{v}_\Omega - \mathbf{v}_{\bar{\Omega}} \right), \tag{A8}$$

where $\Delta p_z = p_\Omega^{(\beta_z)}(z) - p_\Omega^{(0)}(z)$. Since $\beta_z \geq 0$, we have $\Delta p_z \geq 0$, and therefore

$$\|\Delta \mathbf{o}_z\| = \Delta p_z \cdot \|\mathbf{v}_\Omega - \mathbf{v}_{\bar{\Omega}}\|. \tag{A9}$$

By constraining this perturbation to be no larger than τ, we obtain

$$\|\Delta \mathbf{o}_z\| = \Delta p_z \cdot \|\mathbf{v}_\Omega - \mathbf{v}_{\bar{\Omega}}\| \leq \tau, \tag{A10}$$

which proves Eq. 5. □

Closed-form bias limit. Let $D_z = \|\mathbf{v}_\Omega - \mathbf{v}_{\bar{\Omega}}\|$. The drift constraint requires $\Delta p_z \leq \tau / D_z$. For numerical stability, we use a small $\epsilon > 0$, set $p = \text{clip}(p_\Omega^{(0)}(z), \epsilon, 1 - \epsilon)$, and choose

$$\delta_z = \min \left(\frac{\tau}{D_z + \epsilon}, 1 - p - \epsilon \right). \tag{A11}$$

Setting $p_\Omega^{(\beta_z)}(z) = p + \delta_z$ and solving for β_z yields

$$\beta_{\max,z} = \log \frac{(p + \delta_z)(1 - p)}{p(1 - p - \delta_z)}. \tag{A12}$$

In practice, we use this value independently for each localized region and denoising step.

B IMPLEMENTATION DETAILS

B.1 MODEL SETUP

FLUX.2-dev. FLUX.2-dev (Black Forest Labs, 2025) is a 32-billion-parameter flow matching Diffusion Transformer (DiT) model, capable of both text-to-image generation and image editing from single or multiple reference images. Its internal architecture employs DoubleStreamBlock and SingleStreamBlock structures to process visual and textual tokens jointly, and it replaces the traditional CLIP/T5 text encoder combination with a high-parameter LLM for superior prompt comprehension and spatial reasoning. We used `FLUX.2-dev-bnb-4bit`, where the text encoder and DiT were quantized to 4-bit. We ran inference in 28 steps.

FLUX.2-klein. FLUX.2-klein (Black Forest Labs, 2025) is a step-distilled version of FLUX.2. It unifies text-to-image generation, single-reference image editing, and multi-reference composition in a single architecture, delivering state-of-the-art quality with fast inference. It comes in 4B and 9B parameter variants, and we used `FLUX.2-klein-4B`. We ran inference in 4 steps.

LongCat-Image-Edit-Turbo. LongCat-Image-Edit-Turbo (Team et al., 2025) is an image editing model built on top of a DiT-based architecture. It supports a wide range of editing tasks including global editing, local editing, text modification, and reference-guided editing. The Turbo variant is a distilled version of the base LongCat-Image-Edit model, achieving approximately 10× speedup. The model architecture is divided into MM-DiT and Single-DiT blocks, and our method is applied to the Single-DiT blocks. We ran inference in 8 steps.

B.2 FREEID SETUP

Keypoint detection. We use `face_recognition` Python library to detect facial keypoint. Specifically, we detect $K = 4$ semantic facial keypoints: left eye, right eye, nose, and mouth. These keypoints are selected based on cognitive research on face recognition (Valentine, 1991), which emphasizes the importance of internal facial features for identity perception.

Attention masking. For computational efficiency, we extract the facial keypoint mask M_{bias} solely from a single intermediate layer and share it across all layers. Specifically, we use the middle block: Block 10 for FLUX.2-klein and LongCat-Image-Edit-Turbo (out of 20 blocks), and Block 24 for FLUX.2-dev (out of 48 blocks).

Cumulative attention mass. For all main experiments, cumulative attention mass threshold is set to $r = 0.8$.

B.3 CALIBRATION OF THE DRIFT BOUND

The parameter τ specifies the maximum attention output drift permitted by each intervention. For a fixed value of τ, Eq. (5) defines the constraint

$$\|\Delta o_z\| \leq \tau,$$

and Eqs. (A11) and (A12) give the corresponding maximum admissible bias $\beta_{\max,z}$ in closed form. Thus, the derivation determines the bias magnitude for a prescribed drift bound, while the choice of τ controls the overall strength of the intervention.

An architecture aware initial value can be obtained from the number of times the intervention is applied during generation. Let N_{blocks} and N_{steps} denote the number of transformer blocks and inference steps, respectively. Under a first order additive approximation, the cumulative perturbation scale is proportional to

$$\sum_{t=1}^{N_{\text{steps}}} \sum_{b=1}^{N_{\text{blocks}}} \|\Delta o_{b,t}\| \leq N_{\text{blocks}} N_{\text{steps}} \tau.$$

Table A1: Empirically selected drift bounds and estimates obtained from Eq. (A13).

Model	N_{blocks}	N_{steps}	$N_{\text{blocks}}N_{\text{steps}}$	Selected τ	Estimated τ
FLUX.2-klein	20	4	80	2.0×10^{-2}	1.88×10^{-2}
LongCat-Image-Edit-Turbo	20	8	160	1.0×10^{-2}	9.38×10^{-3}
FLUX.2-dev	48	28	1344	1.0×10^{-3}	1.12×10^{-3}

Prompts	
Expression	
1	A photo of the same person laughing with joy.
2	A photo of the same person with a serious expression.
Action	
3	A photo of the same person talking on a phone, outdoor.
4	A photo of the same person drinking coffee in a cafe.
5	A photo of the same person kicking a soccer ball on a green field.
View Angle	
6	A portrait photo of the same person, side profile view.
7	A portrait photo of the same person, three-quarter view.
8	A portrait photo of the same person, bird's-eye view looking down.
Framing	
9	A full body photo of the same person standing on a street.
10	A full body photo of the same person sitting on a park bench.

Table A2: Prompts used in evaluation of image synthesis.

Maintaining a comparable cumulative perturbation budget across architectures therefore gives the inverse scaling rule

$$\tau \approx \frac{1.5}{N_{\text{blocks}}N_{\text{steps}}}. \tag{A13}$$

Table A1 compares this estimate with the values selected empirically for the three evaluated backbones. The estimates closely follow the selected values despite substantial differences in model depth and sampling length.

The additive approximation provides an initialization rule rather than a bound on the final generated sample. Individual perturbations propagate through nonlinear transformer blocks and denoising dynamics, and their effects need not accumulate uniformly. Eq. A13 is therefore used as an architecture aware starting point, after which τ may be adjusted for the desired balance between identity preservation and image quality.

B.4 Evaluation protocol

We evaluate on two datasets, a self-curated Unsplash dataset and CelebA-HQ, each consisting of 25 reference identities. Using 10 prompts per identity, we generate 250 images for each dataset and configuration, and report metrics averaged over the full evaluation set. All images are generated at a resolution of 1024×1024 on a single NVIDIA A100 80GB GPU. Table A2 shows the list of prompts.

B.5 Evaluation metrics

InsightFace. InsightFace (InsightFace AI Technology, 2021) is an open-source face analysis toolkit that provides face detection, alignment, and recognition. It extracts face embeddings using a deep neural network. Face similarity is then computed as the cosine similarity between two such embedding vectors, where faces with similar identity will have embeddings that are closer together in the embedding space. In the context of this paper, InsightFace cosine similarity is computed between the reference image and each generated image to quantify how well the subject's identity has been preserved.

AdaFace. AdaFace (Kim et al., 2022) utilizes loss function that assigns different levels of importance to training samples based on their image quality, approximating quality via feature norms to

form an adaptive margin. This makes AdaFace particularly robust on low-quality images such as those captured under poor lighting or at low resolution. Like InsightFace, it is used here as a face embedding model, and identity preservation is measured as the cosine similarity between reference and generated face embeddings.

CLIP-IQA. CLIP-IQA (Wang et al., 2023) utilizes the pre-trained CLIP model to evaluate image quality and abstract perceptions without requiring task-specific fine-tuning. By adopting an antonym prompt pairing strategy (e.g., "Good photo" vs. "Bad photo") and discarding positional embeddings for input flexibility, it measures the semantic alignment between images and text prompts. This approach effectively captures both technical attributes, such as sharpness, and high-level aesthetic tones. Demonstrating a strong correlation with human judgment, we implement CLIP-IQA via PyIQA [1] using its default configurations.

CLIP-T. CLIP-T (CLIP Text similarity) (Radford et al., 2021) measures the semantic alignment between a generated image and its conditioning text prompt. It embeds both the image and the text into CLIP's shared multimodal embedding space and computes their cosine similarity. It serves as a lightweight proxy for text-image consistency, ensuring that identity-focused attention bias does not cause the model to ignore the prompt during generation.

B.6 Statistical significance

Table A3 summarizes the mean paired differences and their 95% confidence intervals for the results in Table 1.

Table A3: Paired differences and the corresponding 95% confidence intervals of Table 1.

Model	Face Similarity (Δ)		Image Quality	Text Alignment (Δ)
	InsightFace (↑)	AdaFace (↑)	CLIP-IQA (↑)	CLIP-T (↑)
Dataset: Unsplash				
FLUX.2-klein + FreeID	0.059 ± 0.0041	0.058 ± 0.0042	0.00043 ± 0.0015	-0.00016 ± 0.00061
FLUX.2-dev + FreeID	0.040 ± 0.0045	0.040 ± 0.0050	-0.00073 ± 0.0012	-0.00036 ± 0.00040
LongCat-Turbo + FreeID	0.032 ± 0.0031	0.035 ± 0.0033	-0.00018 ± 0.0011	-0.00039 ± 0.00034
Dataset: CelebA-HQ				
FLUX.2-klein + FreeID	0.068 ± 0.0091	0.069 ± 0.0095	0.0011 ± 0.0018	-0.00071 ± 0.00088
FLUX.2-dev + FreeID	0.039 ± 0.0068	0.042 ± 0.0081	0.00015 ± 0.0014	-0.0012 ± 0.00062
LongCat-Turbo + FreeID	0.035 ± 0.0068	0.035 ± 0.0063	-0.0017 ± 0.0011	-0.00033 ± 0.00051

B.7 Settings for generalization experiments

Evaluation. For existing text-to-image personalization and customization methods, we evaluate FreeID on PuLID (Guo et al., 2024) and DreamO (Mou et al., 2025). The evaluation uses the same reference identities and prompt set described in Sec. B.4. These methods already inject reference identity through model-specific learned conditioning mechanisms, providing strong baselines for testing complementary identity gains. For fair comparison, the original model weights, reference images, prompts, denoising settings, and random seeds are kept fixed between each baseline and its FreeID-enhanced variant. The results show that FreeID improves face-similarity metrics on both methods, indicating that our identity-focused intervention can extend beyond native IT2I models without additional training or fine-tuning.

[1] https://github.com/chaofengc/IQA-PyTorch

Prompts used in I2V model evaluation	
Expression	
1	A video of the same person squinting their eyes hard and furrowing their brow against a bright light.
2	A video of the same person puffing out their cheeks and blowing air out slowly.
Action	
3	A video of the same person quickly turning their head to look back over their shoulder.
4	A video of the same person brushing their hair upward, temporarily covering their forehead and eyes with their hand.
5	A video of the same person sliding a pair of large sunglasses down over their face.
View Angle	
6	A portrait video of the same person as the camera quickly pans from a direct side profile to a front view.
7	A portrait video of the same person raising their chin to look directly up at the camera from a low angle.
8	A video of the same person looking down at their phone, captured from a steep high angle.
Framing	
9	A full body video of the same person doing a single, continuous spin in place.
10	A full body video of the same person walking closely past a thick street pillar that momentarily obscures their body and face.

Table A4: Prompts used in evaluation of I2V models with FreeID. The prompts are organized into four categories to assess identity preservation across diverse generation scenarios: expression changes, action sequences, viewpoint variations, and framing/composition changes. Each prompt uses the phrase "the same person" to emphasize identity preservation.

For Image-to-Video (I2V) evaluation, we utilize the prompts listed in Tab. A4, specifically focusing on demanding scenarios where identity preservation is inherently challenging. These cases include significant variations in facial expressions and viewing angles.

C Additional qualitative results

We provide additional qualitative comparisons on FLUX.2-klein (Black Forest Labs, 2025), FLUX.2-dev (Black Forest Labs, 2025), and LongCat-Image-Edit-Turbo (Team et al., 2025) in Fig. A1, A2, A3, A4, and A5.



Figure A1: **Qualitative comparisons on FLUX.2-klein.**



Figure A2: **Qualitative comparisons on FLUX.2-klein.**



Figure A3: **Qualitative comparisons on LongCat-Image-Edit-Turbo.**



Figure A4: **Qualitative comparisons on LongCat-Image-Edit-Turbo.**



Figure A5: **Qualitative comparisons on FLUX.2-dev.**

Table A5: **Comparison with identity-preservation baselines on the Unsplash evaluation set.** All configurations are evaluated using the same 25 reference identities and 10 prompts per identity. The upper two blocks provide controlled same-backbone comparisons on FLUX.2-klein and FLUX.2-dev. The remaining methods use their released native backbones and are shown as system-level context; their results should therefore not be interpreted as isolating the effect of a single method from differences in backbone capability.

Method	Backbone	InsightFace (↑)	CLIP-IQA (↑)	CLIP-T (↑)
Controlled comparison: FLUX.2-klein				
Personalize Anything	FLUX.2-klein	0.244	0.671	0.233
DreamBooth	FLUX.2-klein	0.508	0.542	0.234
FLUX.2-klein + FreeID	FLUX.2-klein	**0.537**	0.656	**0.236**
Controlled comparison: FLUX.2-dev				
Personalize Anything	FLUX.2-dev	0.322	0.610	**0.237**
DreamBooth	FLUX.2-dev	0.588	0.534	0.227
FLUX.2-dev + FreeID	FLUX.2-dev	**0.749**	**0.651**	0.234
System-level comparison with native backbones				
FlashFace	SD1.5	0.300	0.720	0.221
PhotoMaker	SDXL	0.377	0.689	0.221
IP-Adapter + Face	SDXL	0.359	0.597	0.196
InstantID	SDXL	0.759	0.651	0.224
OmniGen	OmniGen	0.484	0.551	0.222
OmniGen2 (Wu et al., 2025b)	OmniGen2	0.394	0.573	**0.237**
PuLID	FLUX.1	0.721	0.637	0.234
DreamO	FLUX.1	0.758	0.688	0.229
DreamBooth	FLUX.1	0.540	0.574	0.211
Diptych Prompting (Shin et al., 2025)	FLUX.1	0.216	0.560	0.232
Personalize Anything	FLUX.1	0.122	0.655	0.226

D Additional experiments

D.1 Comparison with Identity-Preservation Baselines

We further compare FreeID with representative identity-preservation and subject-driven generation methods under the same evaluation protocol used for our main Unsplash experiments (summarized in Sec. 5.2). Specifically, we use the same 25 reference identities and 10 prompts per identity, resulting in 250 generations for each evaluated configuration. All methods are evaluated using the same identity-similarity, image-quality, and text-alignment metrics.

Table A5 reports two types of comparisons. First, where implementation compatibility permits, we provide *same-backbone comparisons* on FLUX.2-klein and FLUX.2-dev. In particular, we compare FreeID with DreamBooth (Ruiz et al., 2023) and Personalize Anything (Feng et al., 2026) using the same FLUX.2 backbone. These comparisons control for backbone capability and therefore more directly isolate the effect of the personalization method. Second, we report representative identity-preservation and unified image-generation systems using their released native backbones. Since these methods differ in backbone architecture, training procedure, and identity-conditioning mechanism, these rows provide *system-level context* rather than a controlled method-level comparison.

The controlled comparisons show that FreeID achieves higher identity similarity than both DreamBooth and Personalize Anything on the same FLUX.2-klein and FLUX.2-dev backbones. For FLUX.2-klein, FreeID obtains an InsightFace similarity of 0.537, compared with 0.508 for DreamBooth and 0.244 for Personalize Anything. The same trend is more pronounced on FLUX.2-dev, where FreeID obtains 0.749, compared with 0.588 and 0.322, respectively. These results indicate that the observed identity-preservation gains cannot be attributed solely to the choice of a stronger FLUX.2 backbone.

The native-backbone comparisons provide a broader view of existing identity-preservation systems. Several methods employ dedicated face encoders, learned identity representations, or other method-

Table A6: **Robustness across reference conditions.** N denotes the number of unique reference identities in each subgroup. Confidence intervals are computed from per-identity paired differences after averaging each metric over the 10 prompts associated with that identity.

Attribute	Group	Method	InsightFace (↑)	CLIP-IQA (↑)	CLIP-T (↑)
Viewpoint	Side view ($N = 19$)	Baseline	0.367	0.598	0.231
		+ **FreeID**	**0.428 (+16.6%)**	0.595 (-0.5%)	0.232 (+0.2%)
		Δ (95% CI)	**+0.0609 [+0.0499,+0.0725]**	-0.0029 [-0.0054,+0.0001]	+0.0005 [-0.0002,+0.0013]
	Front view ($N = 181$)	Baseline	0.485	0.573	0.226
		+ **FreeID**	**0.562 (+15.9%)**	0.571 (-0.3%)	0.226 (-0.1%)
		Δ (95% CI)	**+0.0770 [+0.0720,+0.0820]**	-0.0016 [-0.0025,-0.0007]	-0.0003 [-0.0007,-0.0000]
Face scale	Small ($N = 64$)	Baseline	0.444	0.552	0.226
		+ **FreeID**	**0.521 (+17.3%)**	0.551 (-0.2%)	0.225 (-0.4%)
		Δ (95% CI)	**+0.0769 [+0.0692,+0.0847]**	-0.0010 [-0.0025,+0.0006]	-0.0009 [-0.0014,-0.0003]
	Large ($N = 136$)	Baseline	0.469	0.589	0.227
		+ **FreeID**	**0.544 (+15.9%)**	0.587 (-0.4%)	0.227 (+0.0%)
		Δ (95% CI)	**+0.0747 [+0.0690,+0.0806]**	-0.0021 [-0.0032,-0.0011]	+0.0000 [-0.0003,+0.0004]
Reference resolution	LowRes ($N = 200$)	Baseline	0.236	0.664	0.229
		+ **FreeID**	**0.333 (+41.1%)**	0.657 (-1.1%)	0.227 (-0.9%)
		Δ (95% CI)	**+0.0977 [+0.0902,+0.1050]**	-0.0074 [-0.0085,-0.0063]	-0.0017 [-0.0021,-0.0013]
	HighRes ($N = 200$)	Baseline	0.478	0.585	0.228
		+ **FreeID**	**0.554 (+15.8%)**	0.583 (-0.3%)	0.227 (-0.1%)
		Δ (95% CI)	**+0.0754 [+0.0708,+0.0801]**	-0.0017 [-0.0026,-0.0009]	-0.0003 [-0.0006,+0.0000]

specific conditioning mechanisms, whereas FreeID operates without per-subject optimization or an additionally trained identity-conditioning module. For instance, InstantID attains the highest InsightFace score by relying on a face encoder and ControlNet trained on tens of millions of face images; FLUX.2-dev + FreeID reaches comparable identity similarity (0.749 vs. 0.759) with higher CLIP-T (0.234 vs. 0.224) and no training. Because these configurations use different backbones, we do not interpret their numerical differences as a controlled ranking of the underlying methods. Instead, they provide system-level context for the identity preservation, image quality, and text alignment achieved by FreeID relative to existing approaches.

D.2 Robustness across Reference Conditions and Demographic Groups

We further evaluate the robustness of FreeID on an expanded set of 200 Unsplash reference identities using FLUX.2-klein. For each reference, we use the same 10 prompts, resulting in 2,000 paired baseline–FreeID generations. For statistical analysis, we first average each metric across the 10 prompts for each reference identity and then compute the paired difference between FreeID and the baseline across identities. Thus, N in the following subgroup analyses denotes the number of unique reference identities rather than the number of generated images.

Robustness across reference conditions. We next stratify the expanded evaluation by viewpoint and face scale, and additionally evaluate performance under degraded reference-image resolution. As shown in Table A6, FreeID improves InsightFace similarity in every evaluated condition. The gains are retained for side-view references and small faces, two settings in which identity preservation is particularly challenging. The improvement is also pronounced for low-resolution references: InsightFace increases from 0.236 to 0.333 ($+41.1\%$). For all reference-condition subgroups, the 95% confidence interval of the paired InsightFace improvement remains strictly positive.

Demographic subgroup analysis. We additionally stratify the expanded evaluation by the demographic attributes used in our evaluation. Table A7 reports results across race/ethnicity, gender, and age subgroups. FreeID improves average InsightFace similarity in every evaluated subgroup, and all corresponding 95% confidence intervals for the paired identity improvement remain strictly above zero. The magnitude of the gain varies across subgroups, while changes in CLIP-IQA and CLIP-T remain small in absolute value. These results provide evidence that the identity-preservation improvement is observed across the evaluated demographic subgroups rather than being concentrated in a single subgroup.

Face-keypoint detection failures. Finally, we inspect the external face-keypoint detector on all 200 reference images. We observe four complete detection failures and two additional cases with visibly misaligned keypoints. These observations identify the external keypoint detector as a remaining

Table A7: **Performance across demographic subgroups.** N denotes the number of unique reference identities in each subgroup. Confidence intervals are computed from per-identity paired differences after averaging over the 10 prompts for each identity.

Attribute	Group	Method	InsightFace (↑)	CLIP-IQA (↑)	CLIP-T (↑)
Race/ethnicity	Asian ($N = 62$)	Baseline	0.479	0.560	0.225
		+ **FreeID**	**0.552 (+15.3%)**	0.559 (-0.2%)	0.224 (-0.3%)
		Δ (95% CI)	**+0.0731 [+0.0653,+0.0805]**	-0.0010 [-0.0025,+0.0005]	-0.0006 [-0.0011,-0.0001]
	Black ($N = 43$)	Baseline	0.483	0.576	0.223
		+ **FreeID**	**0.565 (+17.0%)**	0.574 (-0.3%)	0.222 (-0.4%)
		Δ (95% CI)	**+0.0820 [+0.0718,+0.0922]**	-0.0018 [-0.0042,+0.0005]	-0.0008 [-0.0014,-0.0001]
	Other/unspecified ($N = 95$)	Baseline	0.442	0.589	0.230
		+ **FreeID**	**0.516 (+16.7%)**	0.587 (-0.4%)	0.230 (+0.1%)
		Δ (95% CI)	**+0.0740 [+0.0672,+0.0814]**	-0.0022 [-0.0033,-0.0011]	+0.0002 [-0.0003,+0.0006]
Gender	Man ($N = 106$)	Baseline	0.483	0.584	0.226
		+ **FreeID**	**0.552 (+14.2%)**	0.583 (-0.2%)	0.226 (-0.1%)
		Δ (95% CI)	**+0.0686 [+0.0626,+0.0746]**	-0.0013 [-0.0025,-0.0000]	-0.0002 [-0.0006,+0.0002]
	Woman ($N = 94$)	Baseline	0.440	0.571	0.228
		+ **FreeID**	**0.523 (+18.9%)**	0.569 (-0.4%)	0.228 (-0.1%)
		Δ (95% CI)	**+0.0832 [+0.0763,+0.0903]**	-0.0023 [-0.0035,-0.0010]	-0.0003 [-0.0008,+0.0001]
Age	Child ($N = 30$)	Baseline	0.507	0.576	0.220
		+ **FreeID**	**0.590 (+16.3%)**	0.572 (-0.7%)	0.220 (-0.1%)
		Δ (95% CI)	**+0.0825 [+0.0719,+0.0938]**	-0.0041 [-0.0067,-0.0014]	-0.0003 [-0.0012,+0.0006]
	Young ($N = 76$)	Baseline	0.442	0.578	0.233
		+ **FreeID**	**0.517 (+16.9%)**	0.577 (-0.2%)	0.233 (+0.1%)
		Δ (95% CI)	**+0.0748 [+0.0673,+0.0823]**	-0.0013 [-0.0026,-0.0001]	+0.0002 [-0.0002,+0.0006]
	Middle aged ($N = 69$)	Baseline	0.451	0.565	0.227
		+ **FreeID**	**0.521 (+15.4%)**	0.564 (-0.1%)	0.226 (-0.3%)
		Δ (95% CI)	**+0.0695 [+0.0615,+0.0776]**	-0.0007 [-0.0020,+0.0007]	-0.0006 [-0.0012,-0.0001]
	Older ($N = 25$)	Baseline	0.485	0.608	0.222
		+ **FreeID**	**0.570 (+17.6%)**	0.605 (-0.5%)	0.222 (-0.2%)
		Δ (95% CI)	**+0.0853 [+0.0718,+0.1007]**	-0.0032 [-0.0061,-0.0005]	-0.0005 [-0.0013,+0.0002]

failure mode of the current implementation, particularly for reference images in which reliable facial landmarks cannot be obtained.

D.3 Identity preservation in closed-source IT2I models

We evaluate FreeID on open-source IT2I models including FLUX.2-dev (Black Forest Labs, 2025), FLUX.2-klein, and LongCat-Image-Edit-Turbo (Team et al., 2025). We also conduct inference experiments on closed-source IT2I models (e.g., GPT-5.3 by OpenAI, Nanobanana by Google) to assess their identity preservation capabilities.

Celebrity generation. For celebrity subjects, closed-source IT2I models demonstrate superior performance in cross-view generation. We attribute this to the reasoning capabilities of their underlying large language model (LLM) backbone, which leverages extensive knowledge about public figures to infer plausible alternate viewpoints. This LLM-based reasoning enables implicit identity preservation across diverse poses.

Non-celebrity generation. In contrast, for non-celebrity subjects, closed-source models produce visually plausible results but exhibit subtle identity drift. As shown in Fig. A6, GPT-5.3 struggles to preserve fine-grained identity details during cross-view generation of ordinary individuals. The generated side-view images exhibit noticeable identity drift and lack visual fidelity compared to the reference image.

In this setting, FLUX.2-klein + FreeID preserves the reference identity more reliably than the closed-source model. The generated side views more closely resemble the reference subject, suggesting stronger fine-grained identity preservation for non-celebrity subjects.

Implications. These results reveal a critical limitation: closed-source IT2I models alone cannot fully solve the identity preservation problem, particularly for non-celebrity subjects where external knowledge cannot compensate for weak identity grounding. This underscores the importance of explicit identity mechanisms like FreeID, which enhance identity preservation across diverse subject types without relying on external knowledge or reasoning.

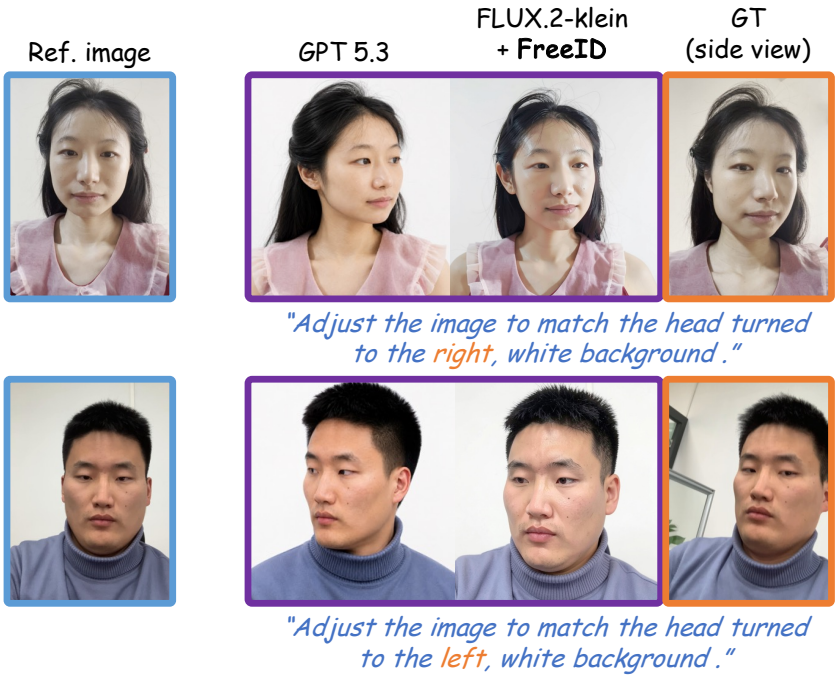


Figure A6: **Cross-view generation on non-celebrity subjects.** Comparison of identity preservation between GPT-5.3 and FLUX.2-klein + FreeID for non-celebrity individuals. While GPT-5.3 produces visually plausible results, it exhibits subtle identity drift and fails to preserve fine-grained facial details during side-view generation, whereas FLUX.2-klein + FreeID maintains robust identity consistency, producing side-view images that closely resemble the reference identity.

D.4 Multi-person personalization

We further provide qualitative results for multi-person personalization in Fig. A7. Unlike the single-reference setting considered in the main experiments, this setting requires preserving multiple identities and assigning each reference identity to the intended generated person. For this experiment, we use FLUX.2-klein and explicitly apply identity-to-person assignment, so that reference identity A is calibrated toward generated person A and reference identity B toward generated person B.

As shown in Fig. A7, the baseline preserves the overall scene and prompt structure, but can mix identity cues across subjects or weaken the resemblance of one of the generated faces. With FreeID, the generated faces better retain identity-relevant facial characteristics for both reference subjects. These examples suggest that the proposed attention calibration can be extended to prompts with multiple reference identities.

D.5 Results for non-face subjects

Our method is designed and evaluated for human faces, where identity differences are subtle and perceptually sensitive. As a preliminary test of whether the underlying insight, that generative models under-attend to identity-critical regions relative to peripheral ones, may also hold beyond faces, we extend our pipeline to non-face subjects by replacing facial landmarks with category-agnostic critical parts, obtained through CLIP/SAM-based localization and DINO feature clustering, and by measuring their contribution to DINO similarity.

Attention analysis on these pseudo-parts reveals the same qualitative tendency observed for faces: peripheral regions receive more attention than identity-critical regions (4.98 vs. 4.75 on average). The gap is smaller than in the face setting, which we attribute to non-face identity cues being distributed across multiple parts rather than concentrated in a small number of landmarks.

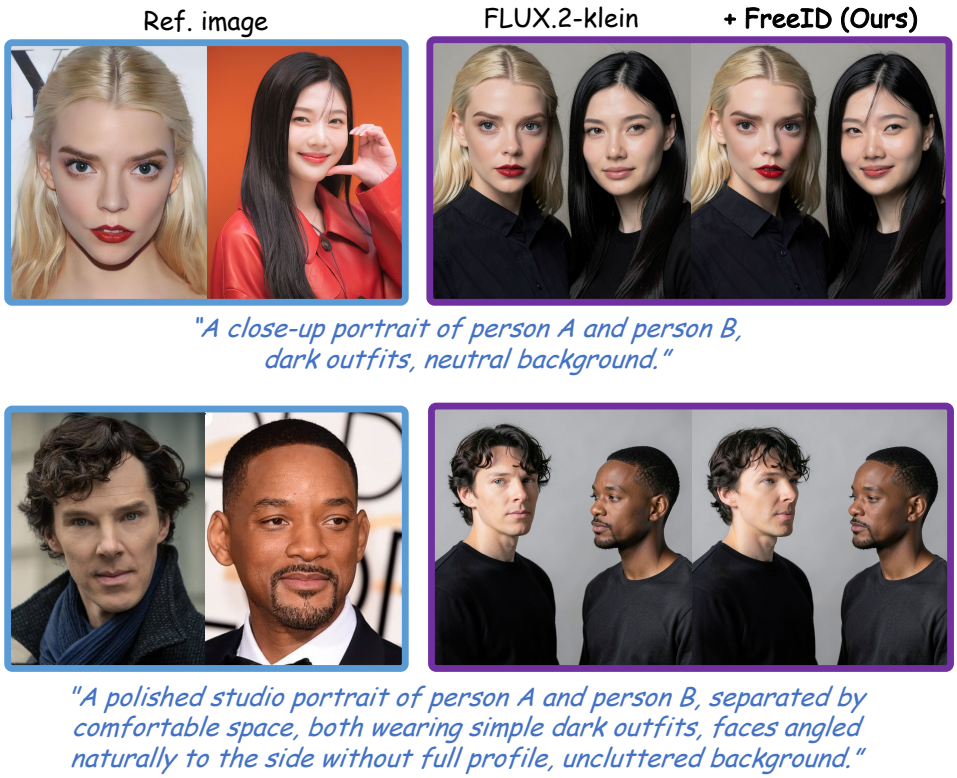


Figure A7: **Multi-person personalization results.** We compare FLUX.2-klein and FLUX.2-klein + FreeID on prompts involving two people. For FreeID, identity-to-person assignment is explicitly applied so that each reference identity is calibrated toward the corresponding generated person. FreeID better preserves identity-relevant facial characteristics for multiple reference subjects, while the baseline can weaken or mix identity cues across generated faces.

We evaluate this extension on three non-face subject types (pets, objects, and characters), using 5 references per type and 5 prompts per reference. As shown in Table A8, applying FreeID with pseudo-parts improves identity preservation (DINO similarity) over the FLUX.2-klein baseline, while maintaining comparable image quality (CLIP-IQA) and text alignment (CLIP-T). These results provide preliminary evidence that this mechanism may extend beyond human faces, which we leave for future work.

Table A8: Quantitative evaluation of FreeID on non-face subjects (pets, objects, characters).

Method	Identity (DINO ↑)	CLIP-IQA ↑	CLIP-T ↑
FLUX.2-klein	0.597	0.687	0.303
+ FreeID	0.620	0.683	0.303

D.6 Ablation study of Otsu's thresholding

Our pipeline binarizes keypoint response maps using Otsu's method to select which tokens receive the attention bias. We motivate this choice by the observation that face scale varies substantially across reference images, from close-up portraits to full-body compositions, and that the appropriate number of selected tokens should vary accordingly rather than follow a fixed rule.

To evaluate the sensitivity of our method to this design choice, we compare Otsu's thresholding against fixed top-percentile thresholds ranging from 1% to 30%, stratified by face scale. On average, Otsu adaptively selects 1.97% of tokens for small faces and 12.20% for large faces, consistent with

the intuition that smaller faces concentrate identity information in fewer tokens. Table A9 shows that no single fixed percentile performs well across both scales: Top-3% is competitive for small faces but underperforms on large faces, while Top-10% performs best for large faces but is suboptimal for small faces. Otsu's thresholding, by adapting to the response-map distribution of each individual reference image, achieves the best overall identity score, outperforming every fixed-percentile baseline while matching or exceeding them on CLIP-IQA and CLIP-T. This confirms that Otsu's thresholding is not an arbitrary design choice but a necessary component for handling the scale variation inherent to in-the-wild reference images.

Table A9: Ablation of Otsu's thresholding against fixed top-percentile thresholds across face scales.

Threshold	Small	Large	Overall	CLIP-IQA ↑	CLIP-T ↑
Otsu	0.461	0.665	**0.584**	**0.652**	**0.234**
Top-1%	0.403	0.592	0.517	0.496	0.232
Top-3%	0.450	0.655	0.573	0.649	0.234
Top-5%	0.424	0.670	0.572	0.644	0.233
Top-10%	0.428	0.675	0.576	0.642	0.233
Top-20%	0.407	0.659	0.558	0.637	0.230
Top-30%	0.413	0.658	0.560	0.630	0.228

D.7 Results for generalization experiments

Fig. A8 shows qualitative results on PuLID (Guo et al., 2024) and DreamO (Mou et al., 2025). Both methods already achieve strong identity preservation, and FreeID further refines fine-grained facial resemblance, such as face shape, the eye region, and local facial details. These examples qualitatively support FreeID's complementary effect on existing personalization and customization pipelines.

Tab. A10 provides the corresponding quantitative comparison on the same Unsplash dataset. FreeID improves both InsightFace and AdaFace scores for PuLID and DreamO, while CLIP-IQA and CLIP-T remain close to the corresponding baseline values. These results indicate that FreeID can be applied beyond the IT2I backbones studied in the main paper and can improve identity preservation on existing personalization and customization pipelines.

Fig. A9 provides qualitative comparisons for image-to-video generation. While the baseline model often loses the subject's identity in later frames, FreeID maintains more consistent identity preservation throughout the temporal sequence.

Table A10: Quantitative comparison on T2I personalization / customization methods. We report model-wise averages on the Unsplash.

Model	Face similarity		Image quality	Text alignment
	InsightFace (↑)	AdaFace (↑)	CLIP-IQA (↑)	CLIP-T (↑)
Dataset: Unsplash				
PuLID	0.721	0.677	0.637	0.234
+ **FreeID**	0.735 (+1.9%)	0.685 (+1.1%)	0.634 (-0.5%)	0.233 (-0.4%)
DreamO	0.758	0.736	0.688	0.229
+ **FreeID**	0.804 (+6.1%)	0.786 (+6.8%)	0.686 (-0.3%)	0.228 (-0.5%)

D.8 Qualitative comparison with token-replacement personalization

We further compare FreeID with Personalize Anything (Feng et al., 2026), a training-free personalization method based on timestep-adaptive token replacement. This qualitative comparison is not intended as a direct performance comparison: Personalize Anything is evaluated with its publicly available FLUX.1-based implementation, whereas FreeID is applied to FLUX.2-klein. Instead, we use this comparison to illustrate the behavioral difference between token replacement and attention calibration.



Figure A8: **Qualitative results on existing personalization and customization methods.** We compare PuLID and DreamO with and without FreeID. While both methods already preserve identity well, FreeID further improves fine-grained facial resemblance. Face-similarity scores are shown below each image.

As shown in Fig. A10, token replacement can strongly preserve the reference appearance, but in cases requiring substantial viewpoint changes, the generated subject may remain overly tied to the reference image. This can lead to copy-paste-like behavior and weaker adherence to prompt-specified changes, such as bird's-eye or side-profile views. In contrast, FreeID does not replace generation tokens with reference tokens. It preserves the original denoising trajectory of the IT2I backbone and selectively strengthens sparse attention links between face-aligned generation queries and identity-relevant reference keys. This allows the model to preserve fine-grained identity cues while still relying on its native generative prior for natural pose, illumination, background, and scene composition.

D.9 User study

Study protocol. We conduct a user study with 114 participants and evaluate perceived identity preservation across three IT2I backbones. We evaluate 15 cases for FLUX.2-klein, 8 cases for FLUX.2-dev, and 15 cases for LongCat. We present illustrative examples from the user study in Fig. A11.

Statistical significance. We assess statistical significance using a two-sided binomial test for each IT2I backbone. Each response is treated as a binary preference between the original backbone and its FreeID-enhanced counterpart, with the null hypothesis assuming equal preference probability, i.e., $\pi = 0.5$. As shown in Table A11, all comparisons yield p-values below 10^{-4}, indicating that the observed user preferences are statistically significant.

Table A11: **User study statistical significance.** Two-sided binomial-test p-values.

Backbone	p-value
FLUX.2-klein	3.66×10^{-276}
FLUX.2-dev	3.48×10^{-85}
LongCat	1.60×10^{-233}

E Additional analysis and visualization

E.1 Adaptive bias selection visualization

Fig. A12 visualizes how the adaptive attention bias changes the generation-side attention map. As shown in Fig. A12(a), the attention is often dominated by peripheral regions such as clothing and hair, while identity-critical facial keypoints such as the eyes, nose, and mouth receive relatively weaker

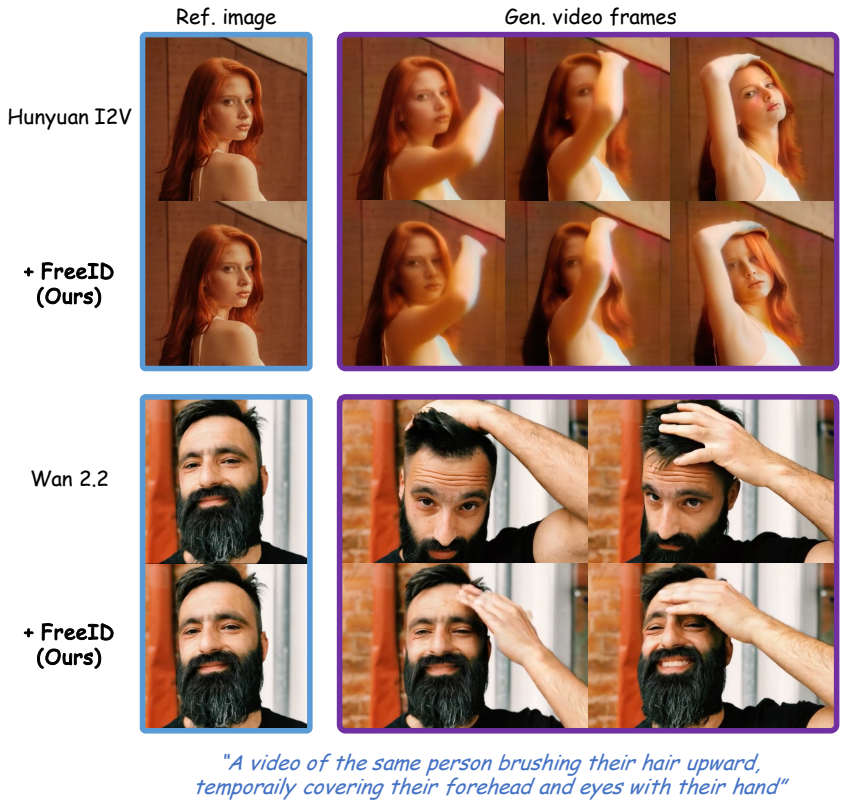


Figure A9: **Qualitative comparison between the Image-to-Video (I2V) model and FreeID.** During dynamic motion, the baseline I2V model often fails to preserve identity consistency with the reference image, whereas FreeID maintains stronger identity preservation throughout the generated video.

attention. This attention imbalance weakens identity preservation, especially for fine-grained facial details.

As shown in Fig. A12(b), FreeID addresses this imbalance by applying adaptive bias only to selected generated tokens associated with identity-relevant facial keypoints. The bias magnitude is adaptively determined for each token by bounding the attention-output drift caused by bias injection. This adaptive bias allows FreeID to strengthen identity-relevant facial attention while limiting excessive perturbation of the attention output, thereby preserving the overall generation behavior of the underlying IT2I backbone.

E.2 Attention pattern analysis

E.2.1 Analysis protocol

We provide additional details on the attention pattern analysis reported in Sec. 3. We evaluate 10 reference images from the Unsplash subset with 5 prompts, resulting in 50 samples in total. For each generated image, we construct three disjoint masks at the token resolution:*face*, *peripheral*, and *background* as visualized in A13. The face region is obtained with InsightFace (InsightFace AI Technology, 2021), peripheral region as the SAM (Kirillov et al., 2023) person mask minus the face, and background as the complement.

Let Z denote the generation tokens and I denote the reference-image tokens. We first compute the attention map $\mathbf{A} = \mathrm{softmax}_{\mathrm{key}}(\mathbf{S})$ over the full key sequence $[T, I, Z]$, and use its reference-token slice $\mathbf{A}^I \in \mathbb{R}^{N_Z \times N_I}$. For each generation token z, we define the *reference-attention mass* and

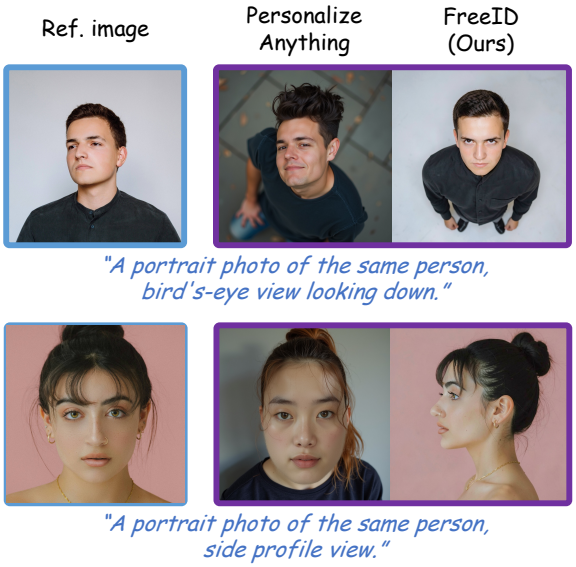


Figure A10: Qualitative comparison between FLUX.1 + Personalize Anything (Feng et al., 2026) and FLUX.2-klein + FreeID. This comparison is not a direct benchmark because the two methods use different FLUX backbones. Personalize Anything can preserve reference appearance strongly, but may remain over-anchored to the reference image, resulting in copy-paste-like behavior and weaker adherence to prompt-specified viewpoint changes. FreeID instead calibrates sparse reference-generation attention links, preserving identity while allowing more natural adaptation to the target scene.

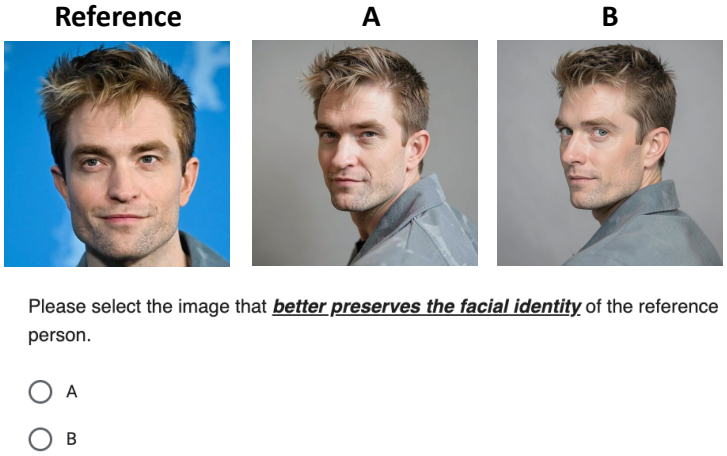
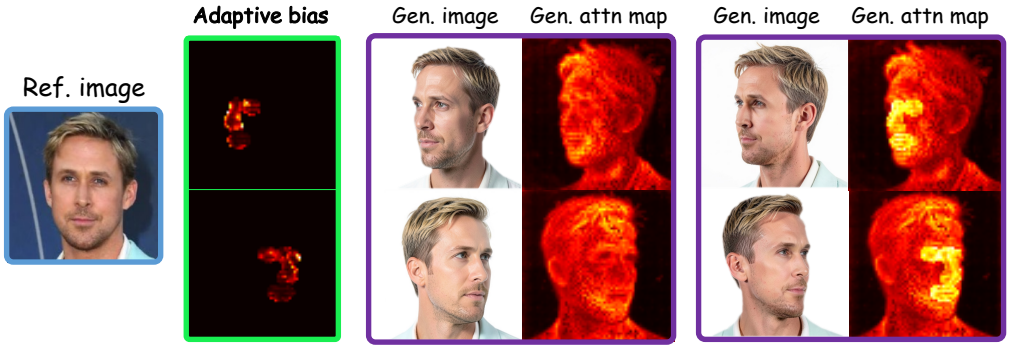


Figure A11: Example images from the user study for facial identity preservation

reference-attention entropy as

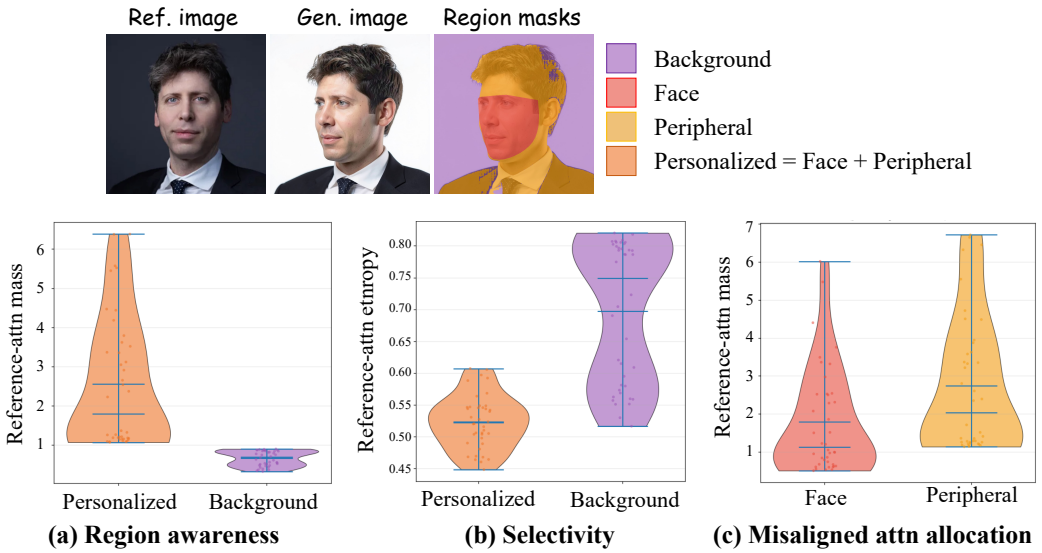
$$A_z = \sum_{i \in I} \mathbf{A}^I_{z,i}, \qquad H_z = -\sum_{i \in I} q_{z,i} \log q_{z,i}, \quad q_{z,i} = \frac{\mathbf{A}^I_{z,i}}{\sum_{j \in I} \mathbf{A}^I_{z,j}}.$$

Here, A_z measures how much of the generation token's attention is assigned to reference tokens, while H_z measures how selectively this reference attention is distributed over reference tokens.



(a) Flux.2-klein (b) **FreeID (ours)**

Figure A12: **Adaptive attention bias visualization.** The base IT2I model often focuses on peripheral regions such as clothing and hair, whereas FreeID selectively strengthens attention to identity-relevant facial keypoints while preserving the underlying generation behavior.



(a) Region awareness **(b) Selectivity** **(c) Misaligned attn allocation**

Figure A13: **Reference-attention patterns across regions.** (a) Personalized tokens draw $\sim 4\times$ more reference-attention mass than background tokens. (b) Personalized tokens also have lower reference-attention entropy, concentrating on a smaller set of reference tokens. (c) Within the personalized region, the peripheral receives more reference attention than the face, revealing a misalignment with identity-defining structure.

Since absolute attention magnitudes vary across samples, we normalize the reference-attention mass and entropy as

$$\tilde{A}_z = \frac{A_z}{\bar{A}}, \qquad \tilde{H}_z = \frac{H_z}{\log N_I},$$

where $\bar{A}$ is the mean of A_z over all generation tokens in the sample.

For each sample s and region ρ, we compute the region-level means $\tilde{A}_\rho^{(s)}$ and $\tilde{H}_\rho^{(s)}$ by averaging over generation tokens inside the region. We then average these region-level values across all samples, yielding the per-sample-normalized, region-averaged quantities.

E.2.2 Analysis results

Across the 50 samples, three consistent patterns emerge, visualized in Fig. A13.

(a) Personalized regions rely more on the reference. The reference-attention mass $\tilde{A}$ is higher in the personalized region than in the background ($\tilde{A}_{\text{pers}} = 2.56$ vs. $\tilde{A}_{\text{bg}} = 0.66$, $\sim 4\times$ ratio). This indicates that personalized tokens draw a much larger share of their attention from reference-image tokens than background tokens do, confirming that personalization is carried predominantly by the reference-attention pathway rather than text- or self-attention.

(b) The reliance is selectively distributed. The reference-attention entropy $\tilde{H}$ is lower in the personalized region than in the background ($\tilde{H}_{\text{pers}} = 0.52$ vs. $\tilde{H}_{\text{bg}} = 0.70$). Beyond merely attending more to the reference, personalized tokens concentrate that attention on a smaller subset of reference tokens, suggesting an explicit token-to-token correspondence between the generated subject and the reference identity.

(c) Within the personalized region, the allocation is misaligned with identity. Decomposing the personalized region into face and peripheral region reveals that the peripheral region receives noticeably more reference attention than the face ($\tilde{A}_{\text{peri}} = 2.73$ vs. $\tilde{A}_{\text{face}} = 1.78$). This is misaligned with identity-defining structure: the face, which carries the strongest identity cues, ends up with the smaller share of reference attention. This imbalance motivates the face-targeted attention reweighting introduced in Sec. 4.

F LIMITATIONS

Inference cost. We measure end to end inference latency on a single NVIDIA A100 80GB GPU using the same generation settings as the main experiments. Each measurement is averaged over 10 timed runs following 3 warmup runs. Table A12 reports latency, computational cost, and peak memory consumption for each backbone.

Across the three backbones, FreeID increases latency by 15.5 percent on average, compared with an average TFLOP increase of 10.0 percent. The difference reflects runtime operations that are not fully represented by FLOP counts, including attention map extraction, token matching and selection, mask construction, and bias application. The difference is small for the two FLUX backbones and larger for LongCat-Image-Edit-Turbo, indicating that the realized overhead also depends on the backbone implementation. Peak memory consumption remains unchanged at the reported precision for all three backbones.

Applicability to other IT2I architectures. FreeID requires reference image tokens to be explicitly present as a distinct subsequence within the unified attention sequence, as in Eq. 1. However, some models encode the reference image via a separate VLM or MLLM before passing an abstract representation to a diffusion decoder (Liu et al., 2025; Z.ai, 2026), leaving no identifiable reference token indices in the DiT attention matrix at generation time. Autoregressive image generation models (Cao et al., 2025) likewise fall outside the DiT framework assumed by FreeID. Extending our approach to these architectures remains future work.

G BROADER IMPACTS

Positive impacts. FreeID improves identity preservation in image-conditioned text-to-image and image-to-video generation without per-subject training or additional model parameters. This can benefit consent-based creative applications, including personalized photo editing, avatar generation, virtual content creation, and controllable media production. Because the method is training-free, it can also reduce the computational cost and repeated data collection required by per-subject fine-tuning approaches.

Potential negative impacts. The same capability may be misused to generate non-consensual identity-preserving images or videos, including impersonation, fake profiles, harassment, fraud, or misleading media. This risk is particularly relevant because FreeID can operate from a single reference image, which may be obtained from publicly available photos. The method may also inherit privacy, fairness, and safety limitations from the underlying generative model and from the face detection or recognition components used to localize identity-relevant regions. For example, identity

Table A12: End to end inference cost on a single NVIDIA A100 GPU.

Model	Latency (s)	TFLOPs	Memory (GB)
FLUX.2-klein	1.32	155.0	21.5
+ **FreeID**	1.53	177.8	21.5
FLUX.2-dev	53.96	8727.6	44.9
+ **FreeID**	60.89	9729.6	44.9
LongCat-Image-Edit-Turbo	4.82	521.8	29.7
+ **FreeID**	5.68	540.9	29.7

preservation quality may vary across image quality, pose, occlusion, lighting, or demographic groups, potentially leading to uneven performance or unintended identity drift.

Responsible use. FreeID should be used only for consent-based personalization and benign creative applications. We encourage future work on stronger identity-consent mechanisms, demographic robustness analysis, and detection/provenance tools for identity-preserving generated media.