Less Is More: Training-Free Acceleration of Identity-Preserved Generation

Anonymous Authors

Submitted to the 4th Workshop on Generative Models for Computer Vision, CVPR 2026

Abstract

Identity-preserved image generation is usually evaluated at a fixed sampling budget, implicitly assuming that more denoising steps improve both visual quality and identity fidelity. We show that this assumption can fail for identity-conditioned Diffusion Transformers. In InfiniteYou, ArcFace identity similarity under the FLUX.1-dev backbone peaks early in the denoising trajectory (ID_{sim} = 0.609 at 4 steps) and then decreases to 0.543 at 28 steps, a drop of 0.066. We refer to this degradation as identity drift*. The trend persists across identity-conditioning scales $\alpha \in [0.25, 1.5]$ and becomes stronger as prompts become more descriptive or style-conflicting, suggesting competition between fixed identity residuals and later text-driven refinements. This observation leads to a simple deployment strategy: replace the 28-step FLUX.1-dev backbone with the distilled 4-step FLUX.1-schnell backbone while keeping InfuseNet frozen. The replacement requires only two configuration changes, uses no retraining, and achieves 5.9× lower latency, +0.028 ID_{sim}, and −0.016 LPIPS. These results highlight denoising-step selection as an overlooked factor in identity preservation and show that distilled few-step backbones can be both faster and more identity-faithful.*

1. Introduction

Identity-preserved image generation aims to synthesize a person across new scenes, outfits, poses, and styles while maintaining recognizable facial identity. This capability is central to personalized content creation, virtual try-on, portrait editing, and avatar generation. Recent methods improve identity alignment by injecting a face representation into a text-to-image model through cross-attention adapters, ControlNet-style branches, or residual side networks. InfiniteYou [1], for example, injects ArcFace [8] embeddings into FLUX [2] Diffusion Transformer (DiT) blocks through InfuseNet, a residual adapter trained with FLUX.1-dev.

Although identity adapters have become increasingly effective, most evaluations treat the denoising schedule as a fixed engineering detail. In practice, InfiniteYou uses FLUX.1-dev for 28 inference steps, which produces high-quality images but costs roughly 10 seconds per image. The common intuition is that additional denoising steps should monotonically improve fidelity. For identity-conditioned generation, however, we find that the opposite can occur: identity similarity is strongest early and degrades as sampling continues.

We call this behavior *identity drift*. In FLUX.1-dev, ArcFace similarity peaks at 4 denoising steps and then drops toward the standard 28-step output. Later steps often refine lighting, background, texture, and style; however, they can also weaken facial geometry and identity-specific cues injected by the fixed identity adapter. This suggests that identity preservation is not only a question of adapter strength, embedding quality, or prompt design—it also depends on where the model stops along the denoising trajectory.

This finding motivates a training-free solution. FLUX.1-schnell is a distilled 4-step FLUX backbone with the same high-level architecture, tokenizers, and text encoders as FLUX.1-dev. Because InfuseNet injects additive residuals rather than replacing backbone features, we test whether the frozen identity adapter transfers directly to FLUX.1-schnell. It does: with only a backbone-path change and `guidance_scale=0`, FLUX.1-schnell becomes a drop-in replacement that is both faster and more identity-preserving.

Prior works on identity-preserved generation [9–11] focus on adapter design and alignment quality, but none analyze how identity fidelity changes along the denoising trajectory. We address this gap.

Contributions.

- We identify *identity drift*: in FLUX.1-dev, ArcFace similarity peaks at 4 steps (0.609) and drops by 0.066 by the 28-step output.
- We show that drift persists across conditioning scales $\alpha \in [0.25, 1.5]$, indicating it is not explained by a single poor choice of adapter strength.
- We show that drift increases with prompt complexity, from 0.008 for minimal prompts to 0.092 for style prompts, while FLUX.1-schnell remains nearly flat ($\leq$0.005).
- We demonstrate a zero-shot backbone replacement: FLUX.1-schnell achieves 5.9× speedup, +0.028 ID_{sim}, and −0.016 LPIPS over 84 samples without retraining.

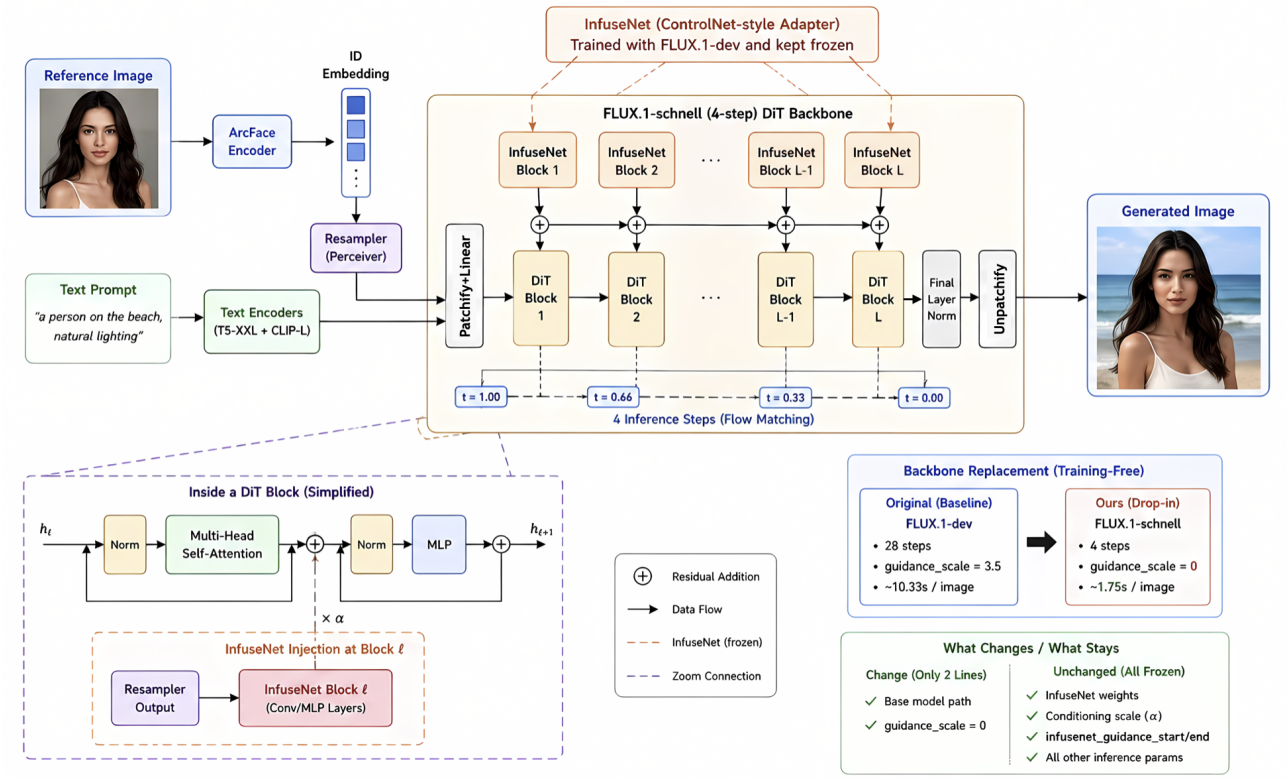

Figure 1. Training-free backbone replacement. The frozen InfuseNet adapter injects ArcFace identity residuals into each DiT block at every denoising step. Replacing FLUX.1-dev with FLUX.1-schnell requires **only two configuration changes**: the backbone path and `guidance_scale=0`. The 4-step backbone naturally operates near the identity-similarity peak observed in FLUX.1-dev, while all adapter weights remain frozen.

2. Background

2.1. FLUX and Flow Matching

FLUX [2] is a family of Diffusion Transformer text-to-image models trained with rectified flow [3, 4]. Rectified flow defines a straight interpolation between data x_0 and noise ε:

$$x_t = (1-t)\,x_0 + t\,\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I), \tag{1}$$

with velocity target $v^* = x_0 - \varepsilon$. This formulation supports large sampling steps after distillation, in contrast to DDPM [5] which requires curved trajectories and 50–1000 steps. FLUX.1-schnell is distilled from FLUX.1-dev via consistency distillation [6, 7] and is commonly used with classifier-free guidance disabled.

2.2. InfiniteYou and InfuseNet

InfiniteYou [1] preserves identity by injecting ArcFace embeddings through InfuseNet. At DiT block ℓ, the hidden state is updated as:

$$h'_\ell = h_\ell + \alpha \cdot \text{InfuseNet}_\ell(\text{id\_emb}), \tag{2}$$

where α controls identity-conditioning strength. The adapter applies an additive residual at each block, so it may remain compatible with a distilled backbone if the hidden-state geometry is sufficiently preserved.

3. Training-Free Backbone Replacement

Our method replaces the FLUX.1-dev backbone in InfiniteYou with FLUX.1-schnell while keeping the identity adapter frozen. The replacement is intentionally minimal: change the backbone path from `FLUX.1-dev` to `FLUX.1-schnell`, and set `guidance_scale=0`. All InfuseNet weights, identity-conditioning scales, prompts, and output resolution remain unchanged.

The intuition has two parts. First, InfuseNet is residual: it nudges backbone features toward the reference identity rather than replacing them. If FLUX.1-schnell preserves a compatible representation after distillation, these residuals should still steer generation toward the target identity. Second, the 4-step trajectory stops near the identity-similarity peak observed in FLUX.1-dev, before later text-driven refinements accumulate enough to dilute the identity signal. Figure 1 summarizes the replacement.

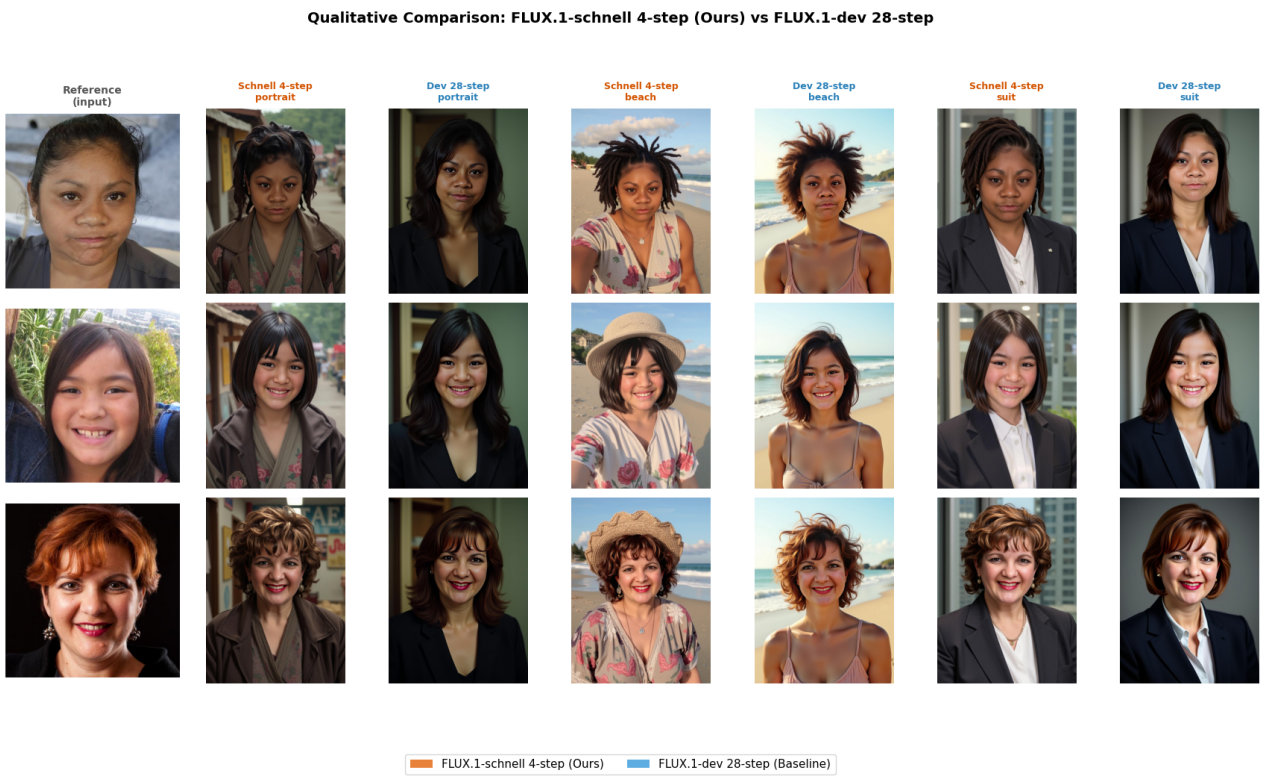

Figure 2. Qualitative comparison across 3 identities and 3 prompts. Column 1: reference face. Columns 2–4: FLUX.1-schnell 4-step (ours) for the prompts related to "portrait", "beach", "suit". Columns 5–7: FLUX.1-dev 28-step (baseline) for matching prompts. FLUX.1-schnell preserves facial structure and identity attributes while reducing latency from 10.29s to 1.73s per image.

4. Experimental Setup

Data. 28 adult identities from FFHQ [15] at 512×512 resolution, filtered for a single visible face, ArcFace confidence >0.7, face width >80px, and estimated age ≥ 18. Each identity is evaluated with three prompts (portrait, beach, professional suit), producing 84 samples per method.

Metrics. ID_{sim}: ArcFace cosine similarity (antelopev2, ↑). LPIPS [14]: perceptual quality, VGG backbone (↓). Latency: wall-clock seconds/image on NVIDIA RTX PRO 6000 Blackwell (96GB VRAM, CUDA 13.0).

Baselines. The main baseline is InfiniteYou with FLUX.1-dev for 28 denoising steps. We also include a FLUX.1-dev 4-step diagnostic on a 10-face portrait subset, included only to separate early stopping from using a distilled backbone.

5. Results

5.1. Qualitative Output Comparison

Figure 2 shows representative qualitative outputs. The 4-step backbone preserves facial structure and identity attributes while maintaining comparable perceptual quality to the 28-step baseline, illustrating that the quantitative gains translate to visible improvements.

Table 1. Main quantitative results (84 samples, 28 identities × 3 prompts, NVIDIA RTX PRO 6000 Blackwell). FLUX.1-schnell improves identity similarity, LPIPS, and latency simultaneously without retraining. The FLUX.1-dev 4-step row is a small diagnostic subset included only to separate early stopping from distilled-backbone effects.

Method	Steps	ID_{sim} ↑	LPIPS ↓	Latency ↓	Speedup	Retrain?
FLUX.1-dev 28-step (baseline)	28	0.5872	0.7253	10.29s	1.0×	—
FLUX.1-dev 4-step (diagnostic)†	4	0.6085	—	1.73s	5.9×	—
FLUX.1-schnell 4-step (ours)	4	**0.6150**	**0.7097**	**1.73s**	**5.9×**	**No**
Δ vs. baseline		+0.028	−0.016	−8.56s		

†FLUX.1-dev 4-step measured on 10-face portrait subset. Not a full-dataset comparison; included to illustrate early-stopping effects separately from backbone distillation.

5.2. Main Quantitative Results

Table 1 shows FLUX.1-schnell improves all reported metrics over the 28-step FLUX.1-dev baseline. The latency reduction is expected, but the identity improvement is not: the 4-step distilled backbone increases ID_{sim} by 0.028 and improves LPIPS by 0.016 without retraining InfuseNet.

Table 2 breaks down ID_{sim} by prompt type. Gains are consistent across portrait, beach, and suit prompts, indicating the improvement is not tied to one scene type. Figure 3 shows per-identity behavior across all 84 samples: FLUX.1-schnell matches or exceeds the FLUX.1-dev 28-step baseline for virtually all identities, confirming the gain is not driven by outliers.

Figure 3 shows per-identity behavior across all 84 samples. FLUX.1-schnell matches or exceeds the FLUX.1-dev 28-step baseline for virtually all identities, confirming the gain is not driven by outliers. The per-prompt averages (dashed lines) further show that the improvement is consistent whether the scene is a portrait, outdoor, or formal setting.

5.3. Identity Drift over Denoising Steps

The per-identity results above show that FLUX.1-schnell consistently outperforms FLUX.1-dev 28-step, but they do not explain *why*. We now analyze the full denoising trajectory to uncover the mechanism.

Figure 5 (left) summarizes the core observation. In FLUX.1-dev, identity similarity is highest early: ID_{sim} reaches 0.609 at 4 steps and decreases to 0.543 by 28 steps (-0.066). We call this degradation *identity drift*.

The result shows that later denoising steps do not uniformly improve all objectives. They may refine visual details and strengthen text alignment while gradually weakening identity-specific structure. This is consistent with Eq. 2: the identity signal is injected as a fixed additive residual, whereas text conditioning continues to shape style, scene, and appearance throughout denoising. As sampling proceeds, prompt-driven refinements can become stronger than the identity residuals, especially when prompts ask for large appearance or style changes. FLUX.1-schnell avoids most of this regime because it operates at the early 4-step endpoint.

Guidance scale insensitivity. Figure 5 (right) shows that ID_{sim} is completely flat across `guidance_scale` $\in \{0.0, 0.5, 1.0, 2.0, 3.5\}$ for FLUX.1-schnell (all values yield $ID_{sim} = 0.591$, variance = 0). Distillation removes classifier-free guidance dependence entirely, eliminating a deployment hyperparameter.

5.4. Conditioning-Scale Ablation

A possible explanation for identity drift is that the identity-conditioning scale α is poorly tuned. To test this, we sweep $\alpha \in \{0.25, 0.5, 0.75, 1.0, 1.25, 1.5\}$. Figure 4 shows that identity drift appears for every α: all curves peak early and degrade by 0.021–0.118 by the 28-step output. Increasing or decreasing adapter strength changes the absolute similarity, but does not remove the drift pattern. This supports the view that drift is a property of the denoising trajectory, not a single bad hyperparameter.

5.5. Prompt Complexity and Text-Identity Competition

We test whether identity drift depends on prompt complexity using 6 prompt types and the top-10 identities by ID_{sim} score. Two trends emerge.

First, drift occurs even for simple prompts, confirming it is not a single-prompt artifact. Second, drift becomes larger as the prompt imposes more appearance or style constraints: minimal prompts produce only a 0.008 drop while style prompts produce a 0.092 drop. This supports the hypothesis that text-identity competition grows when the prompt asks the model to modify appearance more aggressively.

Figure 6 shows the full step curves per prompt type, and Figure 7 summarizes the drift magnitudes and the full 2D similarity landscape side by side.

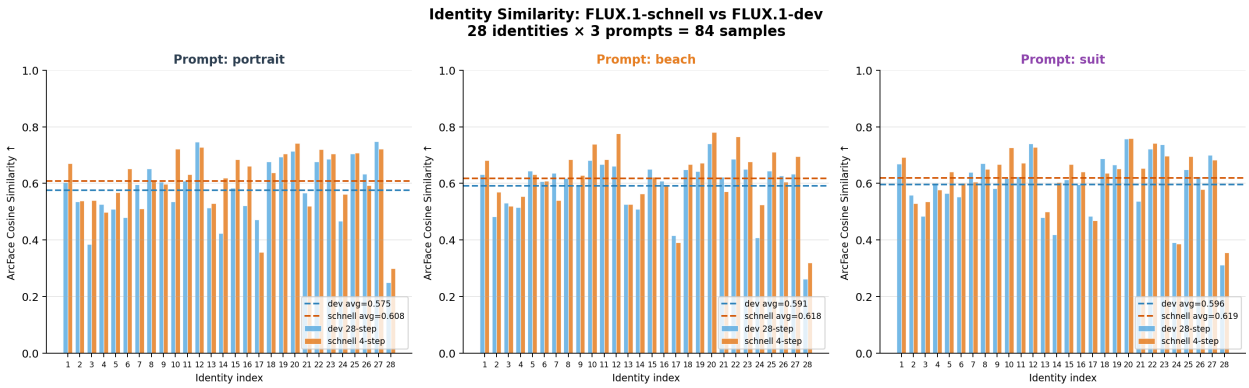

Figure 3. Per-identity ID_{sim} across all 84 samples (28 identities $\times$ 3 prompts). FLUX.1-schnell (orange) matches or exceeds FLUX.1-dev 28-step (blue) for virtually all identities across portrait, beach, and suit prompts. Dashed lines show per-prompt averages. The improvement is *consistent across identities* and not driven by a small number of outlier cases.

Table 2. ID_{sim} by prompt type (28 identities each). Gains are consistent across all three evaluation prompts.

Prompt	FLUX.1-dev 28-step	FLUX.1-schnell 4-step	Δ
Portrait	0.5752	**0.6080**	+0.033
Beach	0.5908	**0.6178**	+0.027
Suit	0.5956	**0.6193**	+0.024
Avg.	0.5872	**0.6150**	+0.028

6. Discussion

Why does the frozen adapter transfer? FLUX.1-dev and FLUX.1-schnell share the same model family, tokenizers, and text encoders. Although distillation changes the weights, the residual identity pathway in InfuseNet only needs the target hidden states to remain geometrically compatible. The successful transfer suggests that the distilled backbone preserves enough feature structure for additive identity residuals to remain useful.

Why not simply use FLUX.1-dev with 4 steps? The 4-step FLUX.1-dev diagnostic suggests that early stopping can improve identity similarity, but it is not a complete solution. It is measured only on a 10-face portrait subset, and a non-distilled backbone is not optimized for high-quality few-step sampling. FLUX.1-schnell is explicitly distilled for 4-step generation, which explains why it combines the identity benefit of early stopping with stable visual quality and low latency.

Implications for identity generation. Identity-preserved generation should not be evaluated only at the default sampling budget. The denoising schedule is part of the identity-control mechanism. For adapter-based systems, identity fidelity may peak before the visually preferred endpoint, so sampling budget, adapter scale, and prompt complexity should be treated as coupled design parameters.

7. Related Work

Identity-preserved generation. IP-Adapter [9] and InstantID [10] inject face embeddings into text-to-image diffusion models. PuLID [11] improves alignment through contrastive objectives. InfiniteYou [1] extends identity preservation to FLUX DiT models through InfuseNet. These works focus on adapter design and alignment quality; our work studies how identity fidelity changes along the denoising trajectory.
Diffusion acceleration. DDPM [5] requires many sampling steps. Progressive distillation [7], consistency models [6], and rectified flow [3, 4] reduce the number of steps. Our work is orthogonal: we show that acceleration can improve identity preservation by avoiding drift.
Adapter transfer. LoRA [12] and model soups [13] show that learned modifications can transfer across related checkpoints. We study a ControlNet-style identity adapter trained with one backbone and deployed zero-shot on its distilled counterpart.

8. Limitations and Responsible Use

Our evaluation uses 28 FFHQ identities and 84 samples, sufficient to expose the drift pattern but not covering full demographic diversity. ArcFace similarity is a useful but imperfect identity metric; future work should combine it with human preference studies and face-region perceptual metrics. FID is not reported as this study is designed around paired identity preservation rather than distribution-level quality.

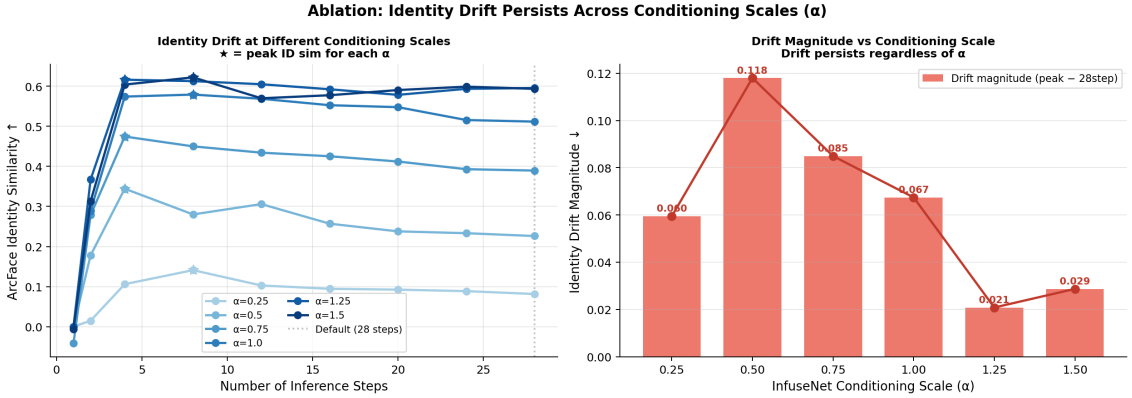

Figure 4. Identity drift persists across conditioning scales α. *Left:* ID_{sim} vs. denoising steps for $\alpha \in [0.25, 1.5]$. All curves peak at 4 steps () and degrade monotonically thereafter. *Right:* Drift magnitude (peak − 28-step value) is positive for all α, ranging 0.021–0.118. Drift is intrinsic to the denoising process, not a scaling artifact.

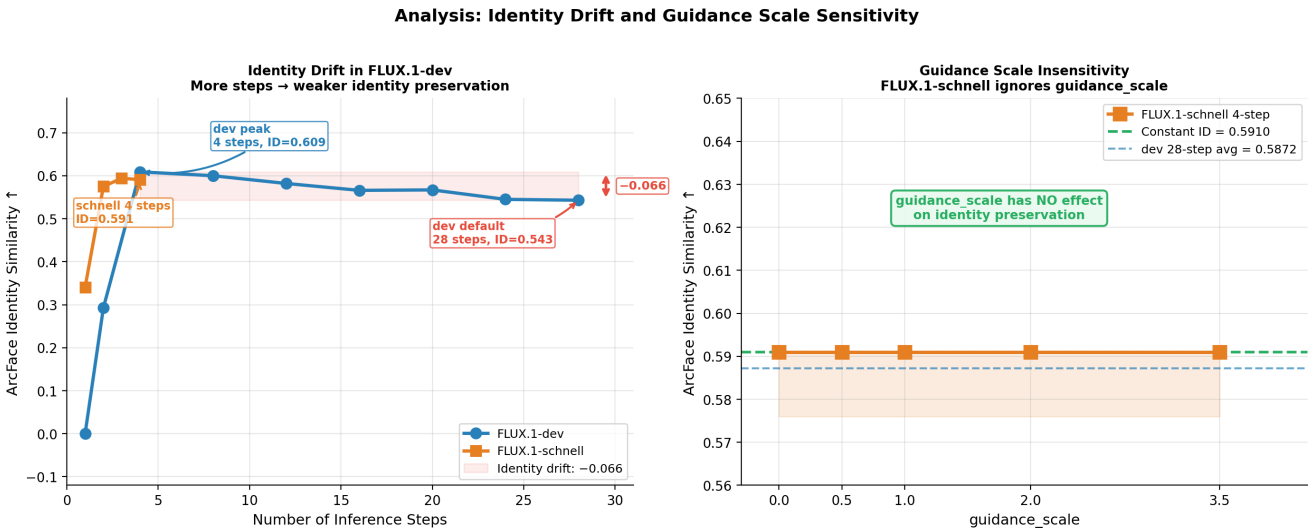

Figure 5. Left: Identity drift in FLUX.1-dev. ID_{sim} peaks at 4 steps (0.609) and degrades monotonically to 0.543 at 28 steps (-0.066 total). FLUX.1-schnell (orange) naturally operates at the peak. **Right: Guidance scale insensitivity.** FLUX.1-schnell's $\text{ID}_{\text{sim}} = 0.591$ is constant across all `guidance_scale` values in $[0.0, 3.5]$ (variance = 0). Users need not tune this parameter.

Because the task involves identity-preserved face generation, responsible deployment requires user consent, provenance tracking, and safeguards against impersonation. Our results should be understood as a technical analysis of sampling dynamics and latency, not as an endorsement of unrestricted face synthesis.

9. Conclusion

We challenge the assumption that more denoising steps always improve identity-preserved generation. In InfiniteYou with FLUX.1-dev, identity similarity peaks at 4 steps (0.609) and drops by 0.066 by the standard 28-step output—a phenomenon we call identity drift. The drift persists across

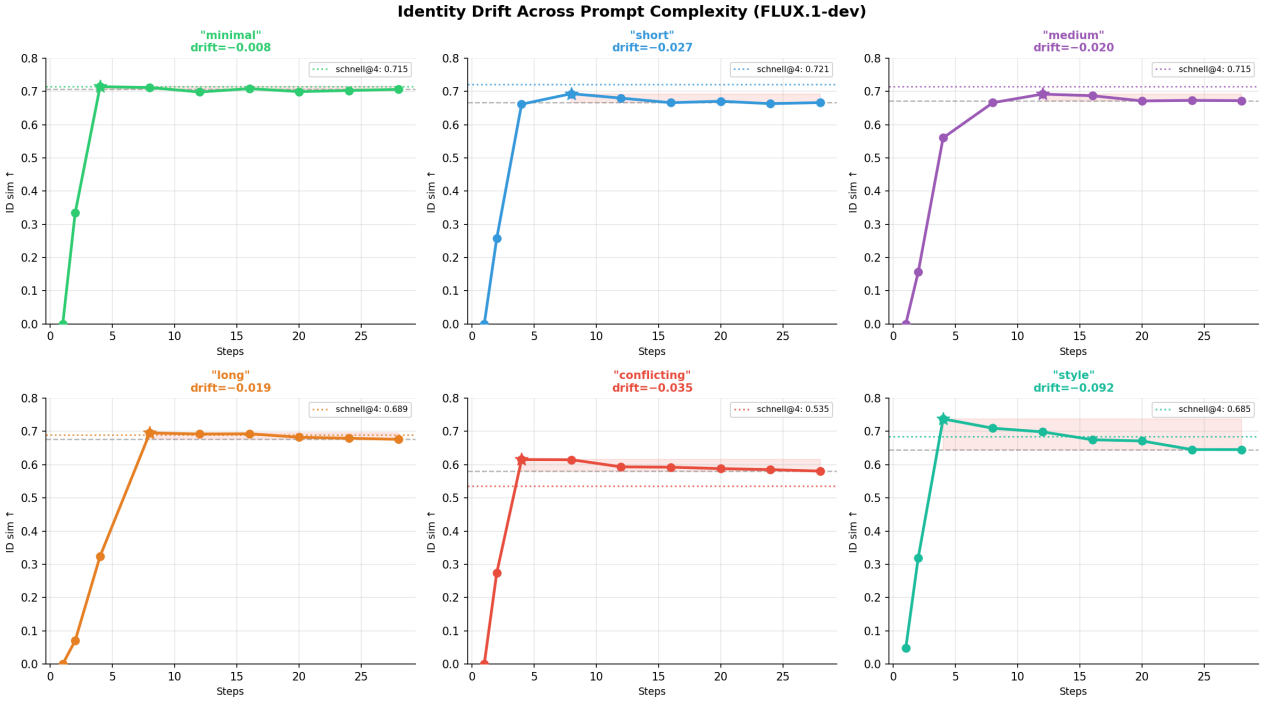

Figure 6. (a) Identity drift across 6 prompt complexity levels (FLUX.1-dev, top-10 identities). Each panel shows ID_{sim} vs. denoising steps for one prompt type. The dotted reference line shows FLUX.1-schnell at 4 steps. Drift is present across *all* prompt types and increases with complexity: -0.008 (minimal) to -0.092 (style). FLUX.1-schnell consistently exceeds or matches FLUX.1-dev's peak.

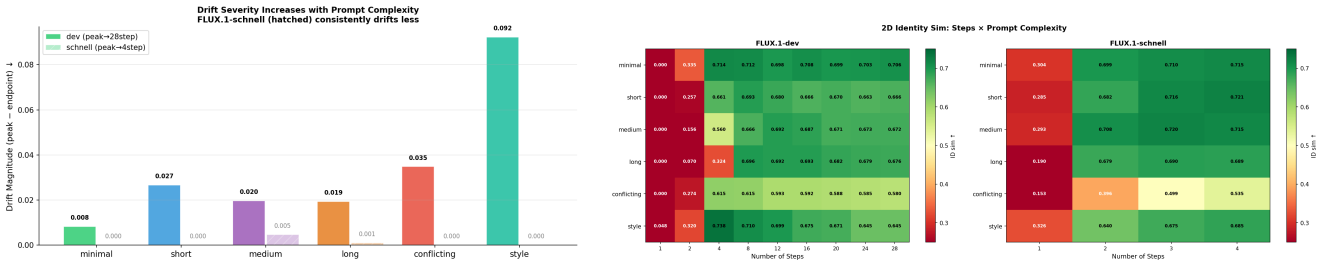

Figure 7. (b) Left: Drift magnitude vs. prompt complexity. Solid bars: FLUX.1-dev drift (peak → 28-step endpoint). Hatched bars: FLUX.1-schnell drift (peak → 4-step endpoint). FLUX.1-dev drift grows from 0.008 (minimal) to 0.092 (style), while FLUX.1-schnell remains $\leq$0.005 across all prompt types. **Right: 2D identity similarity (steps × prompt complexity).** *FLUX.1-dev (left half):* ID_{sim} peaks at steps 4–8 and degrades toward step 28, especially for complex prompts. *FLUX.1-schnell (right half):* ID_{sim} is uniformly high at step 4 across all prompt types, confirming robustness to prompt complexity.

identity-conditioning scales $\alpha \in [0.25, 1.5]$ and grows with prompt complexity (-0.008 for minimal to -0.092 for style prompts), while FLUX.1-schnell remains robust ($\leq$0.005 drift) across all conditions. Motivated by this observation, we replace the 28-step FLUX.1-dev backbone with the distilled 4-step FLUX.1-schnell backbone while keeping InfuseNet frozen. This zero-shot replacement requires only two configuration changes and achieves 5.9× speedup, $+0.028$ ID_{sim}, and -0.016 LPIPS. More broadly, our results show that denoising-step selection is an important and underexplored factor in identity preservation.

References

[1] L. Jiang, Q. Yan, Y. Xiao, et al. InfiniteYou: Flexible Photo Recrafting While Preserving Your Identity. *arXiv:2503.16418*, 2025.

[2] Black Forest Labs. FLUX. `https://github.com/black-forest-labs/flux`, 2024.

[3] X. Liu, C. Gong, and Q. Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *ICLR*, 2023.

[4] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow Matching for Generative Modeling. In *ICLR*, 2023.

[5] J. Ho, A. Jain, and P. Abbeel. Denoising Diffusion Probabilistic Models. In *NeurIPS*, 2020.

[6] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever. Consistency Models. In *ICML*, 2023.

[7] T. Salimans and J. Ho. Progressive Distillation for Fast Sampling of Diffusion Models. In *ICLR*, 2022.

[8] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *CVPR*, 2019.

[9] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. In *ICCV*, 2023.

[10] Q. Wang, X. Bai, H. Wang, Z. Qin, and A. Chen. InstantID: Zero-shot Identity-Preserving Generation in Seconds. In *ECCV*, 2024.

[11] Z. Guo, Y. Wu, Z. Chen, L. Chen, et al. PuLID: Pure and Lightning ID Customization via Contrastive Alignment. In *NeurIPS*, 2024.

[12] E. J. Hu, Y. Shen, P. Wallis, et al. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*, 2022.

[13] M. Wortsman, G. Ilharco, S. Y. Gadre, et al. Model Soups: Averaging Weights of Multiple Fine-tuned Models Improves Accuracy without Increasing Inference Time. In *ICML*, 2022.

[14] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*, 2018.

[15] T. Karras, S. Laine, and T. Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *CVPR*, 2019.