


Identifying and Mitigating Diversity Collapse in Zero-Shot Personalization with I2I Editing Models

Uichan Lee* **Jongeon Baek*** **Sangheum Hwang†**

Department of Data Science

Seoul National University of Science and Technology

{uichan,bjeon,shwang}@ds.seoultech.ac.kr

NeurIPS 2026

∗ Equal contribution

† Corresponding author

This research was supported by Seoul National University of Science and Technology (2025-1371).

Outline

I. Introduction

II. Problem Setup

III. Method

IV. Experiments

V. Analysis

VI. Conclusion

Introduction

Background: Personalized Image Generation

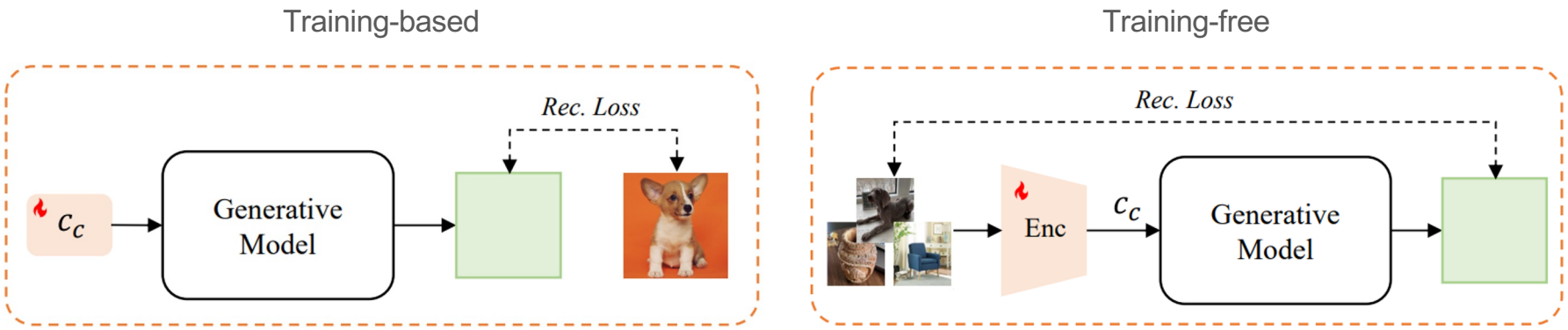
- Recent text-to-image (T2I) diffusion models can generate **high-quality images** from text prompts.
- However, text alone struggles to **reproduce a specific subject consistently**. **Personalization** generates images of a given subject while **preserving its identity**.



Introduction

Existing Methods: Two Approaches to Personalization

- Existing T2I personalization methods are either **training-based** or **training-free**. Training-based methods are costly and tend to overfit, while training-free methods often lose quality.
- So the key challenge is achieving both **identity preservation** and **text alignment** at the same time.



Introduction

Recent Image Editing Models: A Zero-Shot Approach

- **Recent I2I editing models** have demonstrated **strong performance** across a broad range of visual tasks, providing a promising **zero-shot approach** to personalized image generation.



FLUX.1 Kontext: Style transfer (Ghibli)



FLUX.2: Object replacement (cup → cat)

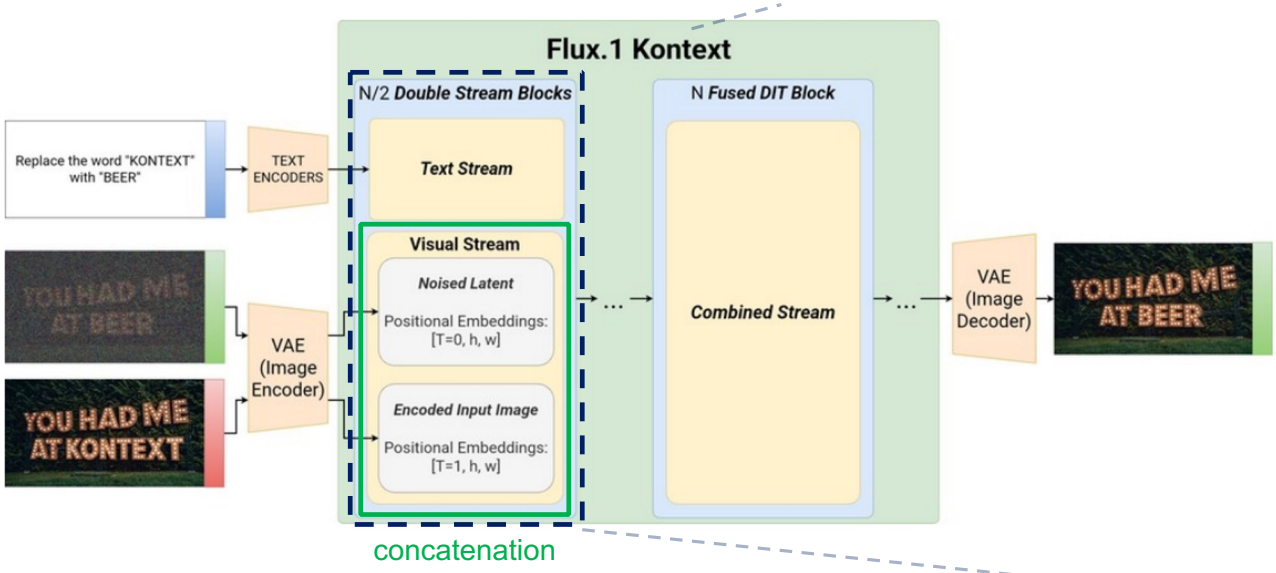
FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space(Black Forest Labs et al., arXiv 2025)
FLUX.2: Frontier Visual Intelligence(Black Forest Labs, 2025)

Problem Setup

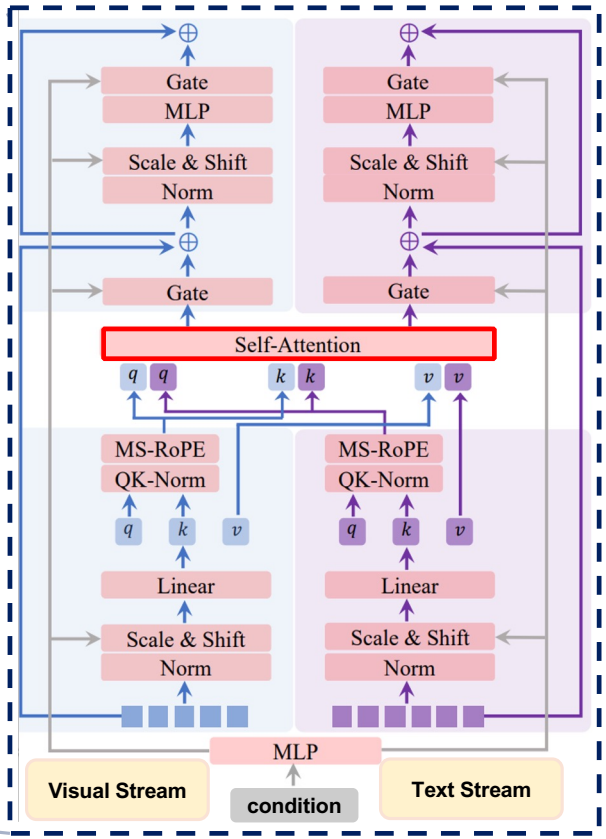
Preliminaries: FLUX I2I Models

- FLUX I2I models condition on a **text prompt** and a **reference image**, concatenating **target latent tokens and reference image tokens** into a single image-side sequence.

[High-level overview of FLUX.1 Kontext]



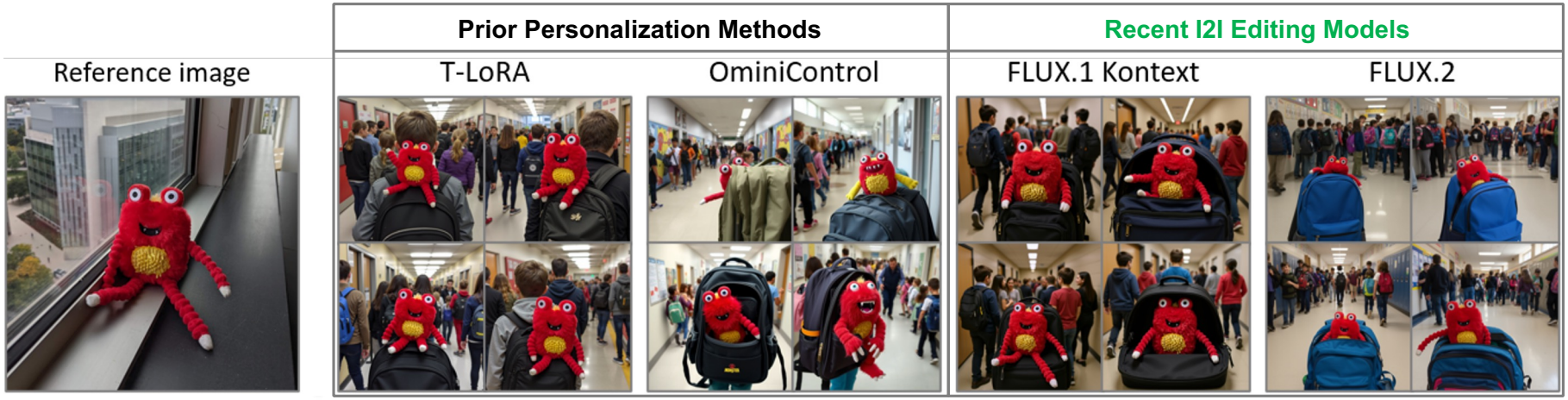
[N/2 Double Stream Blocks]



Problem Setup

I2I Editing Models: Seed-wise Diversity Collapse

- I2I editing models achieve **strong subject identity preservation** and **prompt alignment** in personalization.
- However, they exhibit **seed-wise diversity collapse**: under the same reference image and prompt, different random seeds often produce **highly similar outputs.**

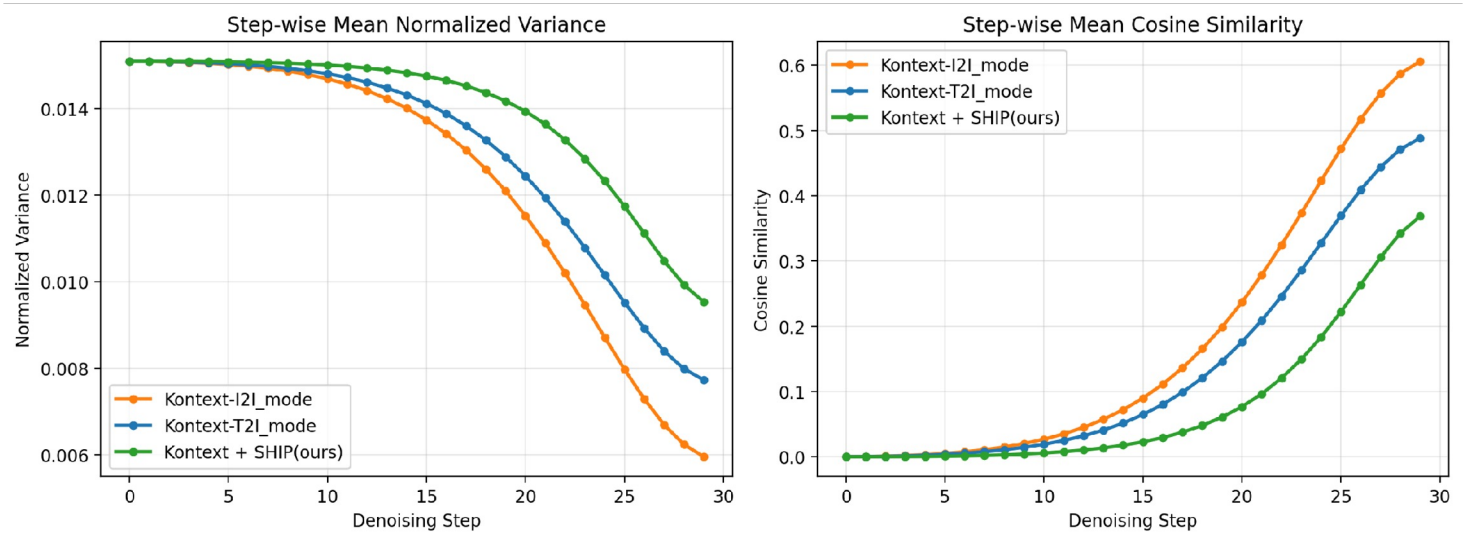


T-LoRA (Soboleva et al., AAAI 2026) · OminiControl (Tan et al., ICCV 2025) · FLUX.1 Kontext (Black Forest Labs et al., arXiv 2025) · FLUX.2 (Black Forest Labs, 2025)

Problem Setup

Motivating Analysis: Seed-wise Trajectory Alignment

- Compared to T2I mode, FLUX.1 Kontext in I2I mode shows **lower normalized variance** and **higher cosine similarity** across random seeds, and the gap grows as denoising progresses.
- This suggests that reference conditioning contributes to **stronger cross-seed trajectory alignment**, which may lead to diversity collapse.

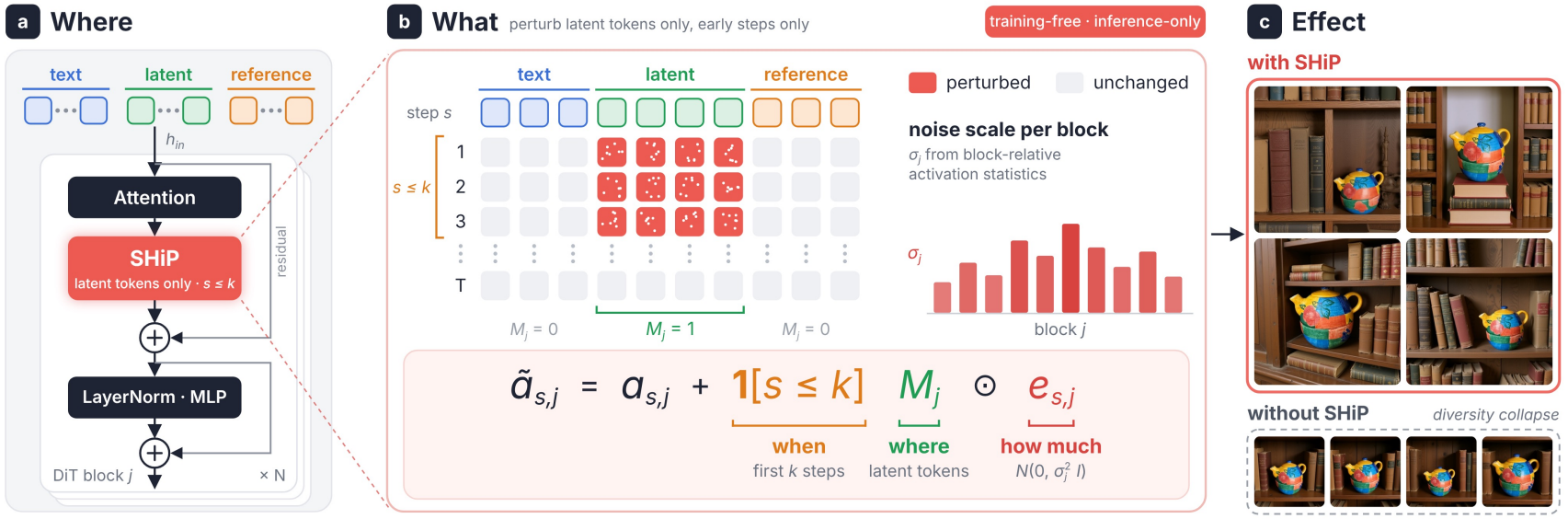


Setup: 6 subjects × 5 prompts × 30 seeds per mode

Method

Selective Hidden Perturbation (SHiP)

- SHiP is a **training-free**, inference-time intervention that injects controlled stochasticity into **target latent tokens only**, keeping text and reference-image tokens unchanged.
- The perturbation is applied to **attention outputs** during the **first k sampling steps**, with noise scaled by block-relative activation statistics.



Experiments

Setup: Benchmark, Baselines, and Metrics

Objective

We evaluate whether SHiP **improves seed-wise diversity** while preserving **identity consistency** and **text alignment**, compared with both training-based and training-free approaches.

Benchmark

Subjects	**15**	DreamBooth Textual Inversion
Prompts	**9**	DreamBench++
Seeds	**30**	per pair
Total	**4,050**	images per method

Baselines

T2I Training-based	DreamBooth, T-LoRA
T2I Training-free	IP-Adapter
Image Editing	OminiControl, FLUX.1 Kontext, FLUX.2
Ours	**FLUX.1 Kontext + SHiP** **FLUX.2 + SHiP**

Metrics

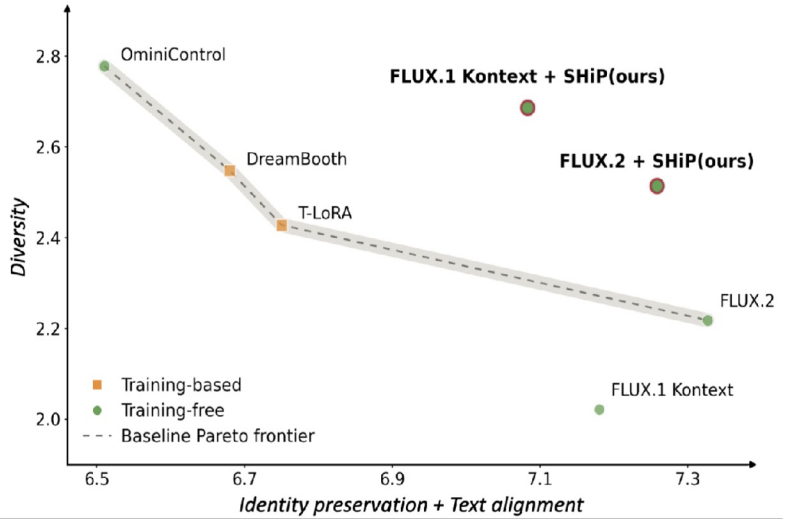
DreamBench++ protocol	GPT-5-mini 0–4 scale
Identity Preservation	**GPT-ID ↑**
Text Alignment	**GPT-Text ↑**
Diversity	**Vendi ↑ / LPIPS ↑**

Experiments

Quantitative Results: Fidelity–Diversity Trade-off

- FLUX.1 Kontext and FLUX.2 achieve **strong identity consistency** and **text alignment**, but exhibit **substantially lower diversity** than other baselines.
- SHiP **significantly improves diversity** with comparable GPT-ID and GPT-Text, moving beyond the baseline **Pareto frontier.**

Method	GPT-ID ↑	GPT-Text ↑	Vendi ↑	LPIPS ↑
DreamBooth [25]	3.35	3.33	1.965	0.583
T-LoRA [27]	3.44	3.30	1.865	0.563
IP-Adapter [33]	2.44	2.60	1.965	0.570
OminiControl [29]	2.87	3.64	2.133	0.645
FLUX.1 Kontext [11]	3.48	3.70	1.547	0.475
+ SHiP (Ours)	3.43 (−0.05)	3.66 (−0.04)	2.066 (+0.519)	0.621 (+0.146)
FLUX.2 [2]	3.40	3.93	1.699	0.519
+ SHiP (Ours)	3.33 (−0.07)	3.93 (+0.00)	1.909 (+0.210)	0.605 (+0.086)



Experiments

Qualitative Results: Diversity without Losing Fidelity

- T-LoRA shows **limited prompt alignment**, and OminiControl shows weaker identity consistency.
- SHiP **retains the fidelity** of FLUX models while introducing **variations in background, pose, and scene composition.**



Experiments

Beyond FLUX: Generalization to Other I2I Editing Models

- Diversity **collapse also appears** in Qwen-Image-Edit and FireRed-Image-Edit.
- SHiP **generalizes** to these models, **improving diversity** while keeping identity and text alignment nearly unchanged.

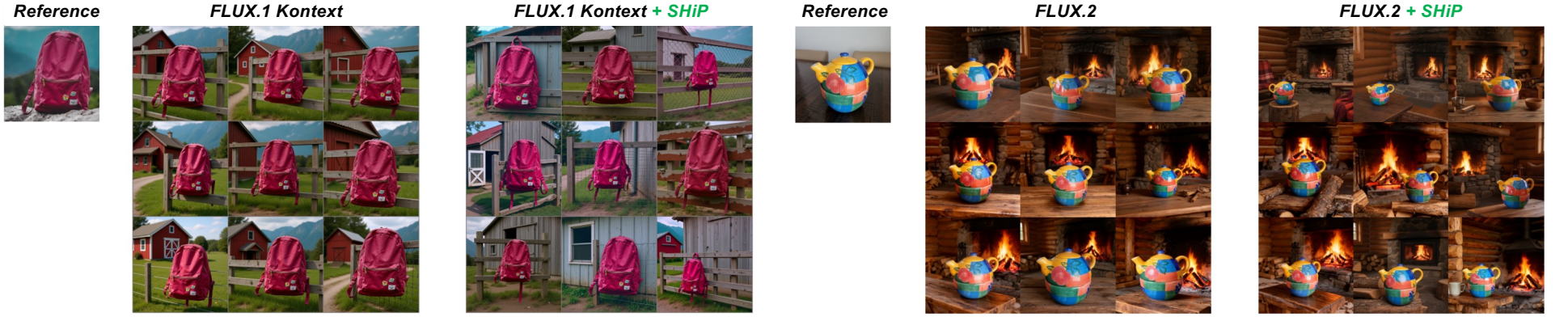


*A photo of the **cat** intently watching birds from behind a garden bush*

Method	GPT-ID ↑	GPT-Text ↑	Vendi ↑	LPIPS ↑
Qwen-Image-Edit [31]	3.37	3.89	1.670	0.548
+ SHiP (Ours)	3.35 (−0.02)	3.88 (−0.01)	1.927 (+0.257)	0.624 (+0.076)
FireRed-Image-Edit [30]	3.37	3.90	1.553	0.484
+ SHiP (Ours)	3.38 (+0.01)	3.90 (+0.00)	1.781 (+0.228)	0.592 (+0.108)

Experiments

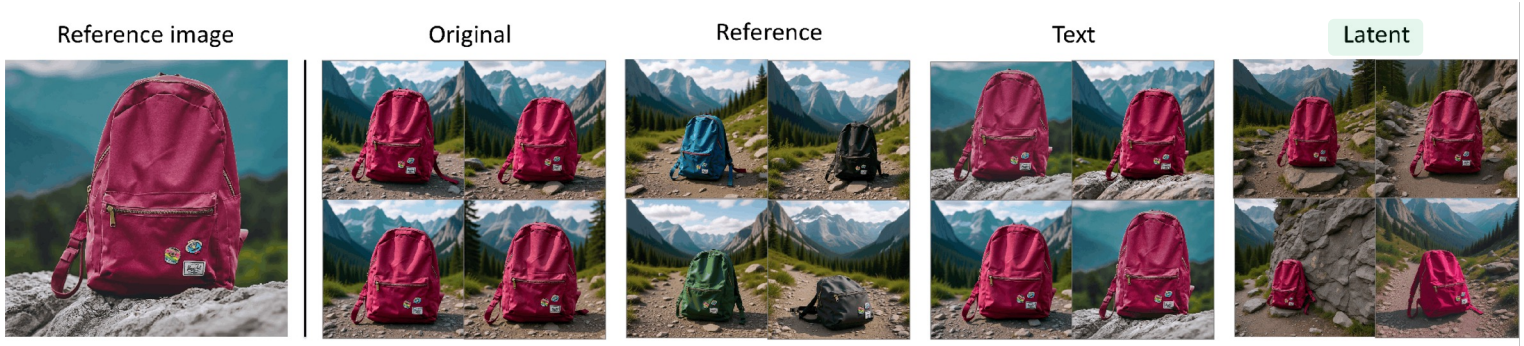
Additional Qualitative Results Across Four I2I Editing Models



Analysis

Ablation: Perturbation Target Tokens

- Perturbing **reference tokens** improves diversity, but mainly through **identity drift.**
- Perturbing **latent tokens** achieves the **best trade-off**, improving diversity while preserving identity and text alignment.



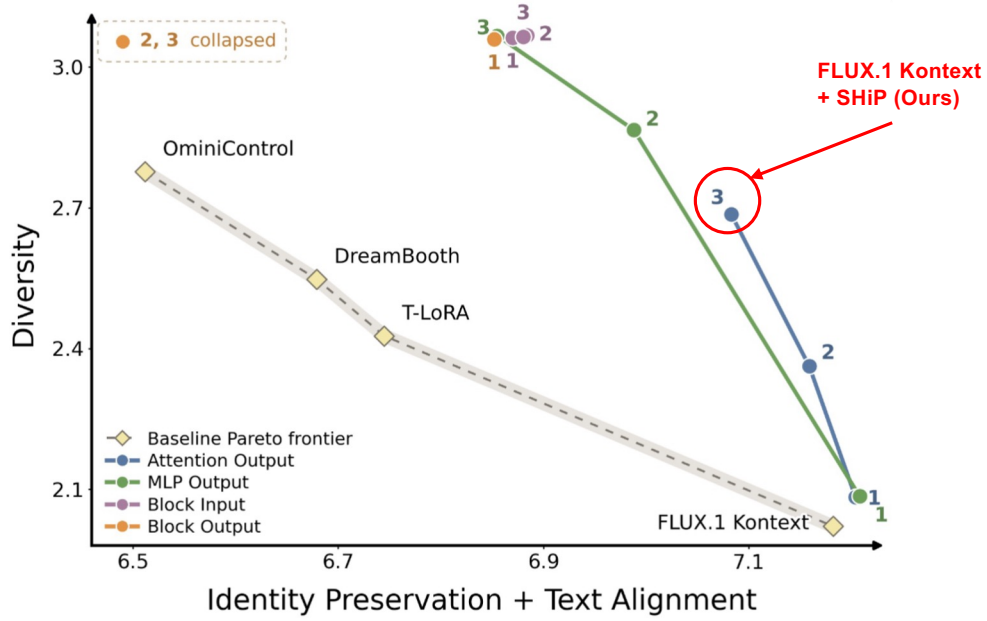
*A photograph of the **backpack** left on a rocky trail in the mountains*

Token	ID	Text	Vendi	LPIPS
Ref.	2.86	**3.71**	1.820	0.521
Text	**3.60**	3.43	1.611	0.495
Latent	3.43	3.66	**2.066**	**0.621**

Analysis

Ablation: Perturbation Sites and Strength λ

- Perturbation at **any site improves diversity**, but the quality–diversity trade-off **varies by site.**
- **Attention-output perturbation** scales most stably with λ and achieves the best fidelity–diversity trade-off.

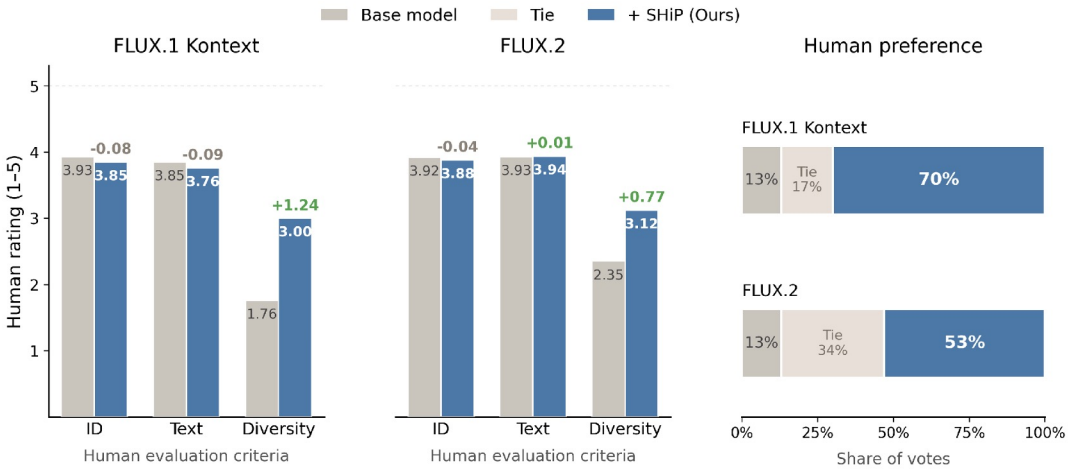


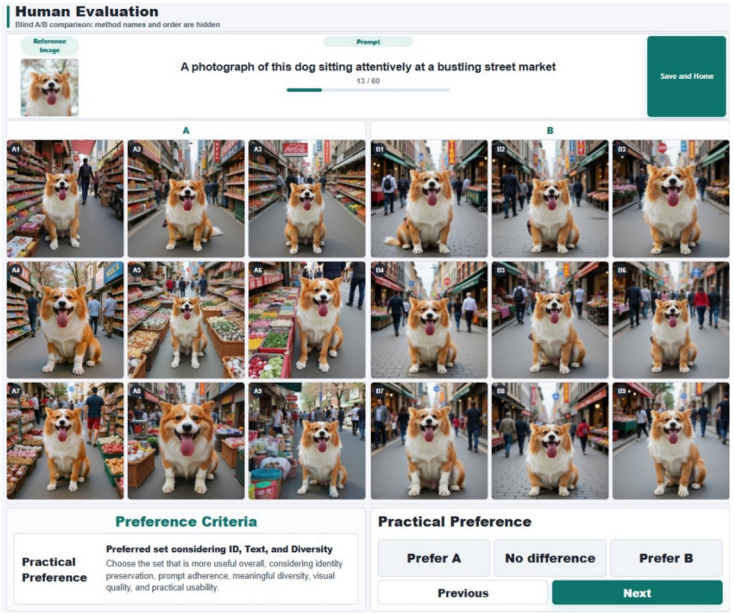
- Block Output: collapses at λ = 2, 3
- Block Input: limited control
- MLP Output: fidelity drops sharply
- **Attention Output: stable gain → used in SHiP (λ = 3)**

Analysis

Human Evaluation: Meaningful Diversity and Preference

- Human ratings show SHiP **greatly improves diversity** with **minimal loss** in ID and text alignment.
- Participants **preferred SHiP outputs** in blind A/B comparisons (**70% / 53%** vs. 13%).





Conclusion

Summary and Contributions

- We identify diversity collapse in I2I editing-based personalization and propose SHiP to mitigate it.

Our Contributions

1. Problem

I2I editing models preserve ID and text well, but collapse to nearly identical outputs across seeds.

2. Method

SHiP perturbs only latent-token attention outputs in early steps, with no training.

3. Results

Higher diversity across four models with minimal fidelity loss, preferred in a blinded human study.

Limitation	The underlying mechanism of diversity collapse remains to be fully understood, motivating future work on broader diversity strategies.