COPYCAT: IMPROVING FINE-GRAINED SUBJECT CONSISTENCY IN SUBJECT-TO-IMAGE MODELS WITHIN SECONDS

Anonymous authors
Paper under double-blind review

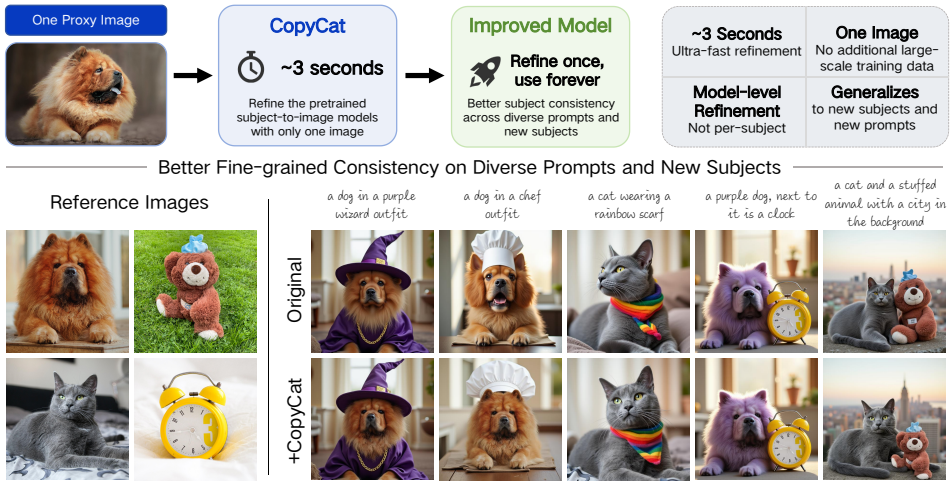


Figure 1: **Overview of CopyCat.** Using a single proxy image, CopyCat learns a lightweight Fine-grained Consistency LoRA to perform a one-time, few-second refinement of a pretrained subject-to-image model. The refined model is directly applicable to unseen reference subjects and diverse prompts without further subject-specific optimization. Representative examples demonstrate improved preservation of fine-grained subject details while maintaining prompt fidelity.

ABSTRACT

Recent subject-to-image models have achieved impressive progress in personalized image generation, yet they still struggle to preserve fine-grained subject-specific details. A major reason is the lack of high-quality fine-grained identity supervision: real paired data are expensive to collect, while synthesized training pairs often preserve only coarse subject appearance and fail to capture subtle subject-specific details. In this work, we propose CopyCat, a lightweight model-refinement framework that improves fine-grained subject consistency within only a few seconds. CopyCat performs a one-time refinement of a pretrained subject-to-image model by attaching a lightweight Fine-grained Consistency LoRA (FCLoRA) and optimizing it using a single proxy image, which is used as both the conditioning image and the reconstruction target. This exact self-reconstruction objective substantially simplifies the optimization task, enabling effective fine-grained refinement within only a few seconds. The refinement is performed only once; the resulting model can be directly applied to diverse unseen reference subjects and prompts without further subject-specific optimization. We further revisit subject-to-image LoRA training in double-stream diffusion transformers and find that removing dedicated text-stream LoRA while retaining image-stream and shared-block LoRA improves subject consistency. Extensive experiments on DreamBench and XVerseBench demonstrate consistent improvements in fine-grained subject consistency across representative subject-to-image models under both single- and multi-subject settings.

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.

1 INTRODUCTION

Recent advances in text-to-image generation have enabled remarkable progress in high-quality visual content synthesis. Building upon these powerful generative models, subject-driven image generation aims to synthesize novel images that not only follow textual instructions but also faithfully preserve the identity of one or more reference subjects, making it valuable for applications such as personalized content creation, virtual photography, and digital entertainment. Early personalization approaches, including DreamBooth (Ruiz et al., 2023) and Textual Inversion (Gal et al., 2022), optimize the model for each target subject, while more recent subject-to-image (S2I) models, such as UNO (Wu et al., 2025c), enable feed-forward generation by learning subject-conditioned generation from large-scale training data. Despite these advances, existing subject-to-image models still struggle to faithfully preserve fine-grained subject-specific details, limiting the fidelity of personalized image generation.

A major obstacle lies in the supervision itself. Existing subject-to-image models are primarily trained using large-scale synthesized subject-image pairs because collecting real paired data with accurate identity correspondence is prohibitively expensive. Although synthesized data provide effective supervision for coarse subject appearance, they often fail to preserve subtle subject-specific details. Consequently, current models often preserve the coarse identity of the reference subject while failing to faithfully reproduce the fine-grained details that distinguish one subject from another.

One straightforward solution is to improve the quality and scale of synthesized training data. However, this requires repeatedly constructing large-scale datasets and retraining expensive subject-to-image models, making it difficult to quickly benefit existing pretrained models. Instead, we ask a different question: *can an existing pretrained subject-to-image model recover its missing fine-grained subject consistency through an extremely lightweight refinement without requiring additional large-scale training data?*

Motivated by this question, we propose **CopyCat**, a lightweight model-refinement framework that significantly improves fine-grained subject consistency within only a few seconds. As illustrated in Figure 1, CopyCat performs a one-time refinement of an existing subject-to-image model by introducing a lightweight *Fine-grained Consistency LoRA* (FCLoRA). During refinement, a single image, referred to as a *proxy image*, is used as both the conditioning image and the reconstruction target, forming an exact self-reconstruction objective. In the comparison in Figure 6(a), using the identical image as the reconstruction target improves subject consistency, whereas using another image of the same subject brings no improvement. Because the conditioning image and the reconstruction target are perfectly aligned, the refinement objective eliminates the need to model subject variations, allowing the lightweight LoRA to focus exclusively on strengthening fine-grained visual correspondence. We interpret the transfer to unseen subjects as strengthening fine-grained consistency capabilities already present in the pretrained model. This refinement is performed only once. The resulting model can subsequently be applied to diverse unseen reference subjects and prompts without subject-specific optimization.

Beyond the proposed model refinement, we further revisit the design of subject-to-image LoRA in existing double-stream diffusion transformers. Removing LoRA from the dedicated text stream while retaining it in the image stream and shared Single-Stream Blocks consistently improves subject consistency in our experiments while reducing the number of trainable parameters. This simple design provides a stronger subject-to-image LoRA and serves as an independent improvement that can be readily adopted during subject-to-image model training.

Our contributions are summarized as follows:

- We propose **CopyCat**, a lightweight model-refinement framework that substantially improves fine-grained subject consistency through only a few seconds of one-time refinement.
- We identify exact self-reconstruction using a single proxy image as the key to effective few-second model refinement. This finding provides a simple yet effective paradigm for improving fine-grained subject consistency without requiring additional large-scale training data.

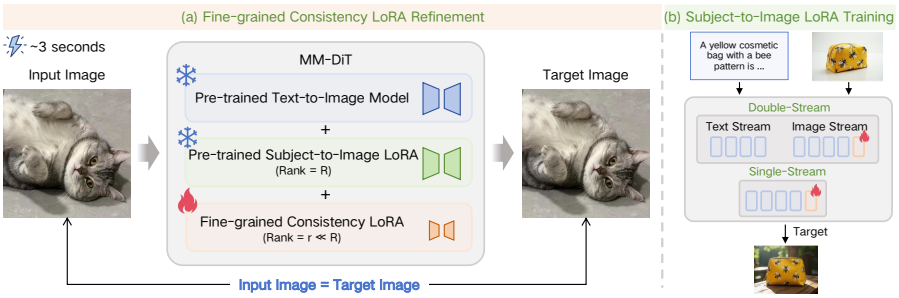


Figure 2: **Overview of the proposed method.** (a) CopyCat attaches a lightweight Fine-grained Consistency LoRA (FCLoRA) to a frozen subject-to-image model and performs a one-time refinement completed within only a few seconds using a single proxy image, where the conditioning image and the reconstruction target are pixel-wise identical. (b) We revisit the design of Subject-to-Image LoRA in double-stream diffusion transformers and remove LoRA from the dedicated text stream while retaining it in the image stream and the subsequent shared Single-Stream Blocks.

- We revisit the LoRA design in double-stream subject-to-image models and demonstrate that removing dedicated text-stream LoRA while retaining image-stream and shared-block LoRA outperforms the conventional full LoRA configuration.

2 Related Work

2.1 Optimization-based Subject Personalization

Early subject personalization methods customize a pretrained text-to-image model through subject-specific optimization. Textual Inversion (Gal et al., 2022) learns customized textual embeddings while keeping the diffusion model frozen, whereas DreamBooth (Ruiz et al., 2023) fine-tunes the model using only a few reference images with prior preservation. HyperDreamBooth (Ruiz et al., 2024), Custom Diffusion (Kumari et al., 2023), Perfusion (Tewel et al., 2023), and LoRA-based adaptation (Hu et al., 2022) further improve personalization efficiency by reducing the optimization cost or the number of trainable parameters. SISO (Shpitzer et al., 2025) optimizes LoRA parameters for a given target subject using similarity losses on generated images, and reuses the optimized model across prompts for that subject. To support multi-concept personalization, subsequent studies explore attention manipulation, concept decomposition, and parameter composition, including Break-A-Scene (Avrahami et al., 2023), Cones (Liu et al., 2023a;b), Mix-of-Show (Gu et al., 2023), Concept Weaver (Kwon et al., 2024), ZipLoRA (Shah et al., 2024), Orthogonal Adaptation (Po et al., 2024), CLoRA (Meral et al., 2025), ConceptSplit (Lim et al., 2025), and FreeLoRA (Zheng et al., 2025). CopyCat instead optimizes identical-image reconstruction on a proxy subject and reuses the refined model across different subjects without further subject-specific optimization.

2.2 Feed-forward Subject-to-Image Generation

Feed-forward methods eliminate subject-specific optimization by directly extracting subject representations from one or more reference images. Representative approaches include ELITE (Wei et al., 2023), BLIP-Diffusion (Li et al., 2023), SuTI (Chen et al., 2023b), IP-Adapter (Ye et al., 2023), PhotoVerse (Chen et al., 2023a), Taming Encoder (Jia et al., 2023), PhotoMaker (Li et al., 2024), InstantBooth (Shi et al., 2024), InstantID (Wang et al., 2024b), RealCustom++ (Mao et al., 2025), and PuLID (Guo et al., 2024). CustomContrast (Chen et al., 2025b) addresses entanglement between subject identity and pose, view, or background through contrastive learning. To support multiple customized subjects, Subject Diffusion (Ma et al., 2024), FastComposer (Xiao et al., 2025a), MS-Diffusion (Wang et al., 2025), MIP-Adapter (Huang et al., 2025), AnyStory (He et al., 2025), and MoA (Wang et al., 2024a) improve subject disentanglement through subject-aware conditioning, regional attention, or unified modulation. More recently, diffusion-transformer-based frameworks, including OminiControl (Tan et al., 2025), OmniGen (Xiao et al., 2025b), OmniGen2 (Wu et al., 2025a), DreamO (Mou et al., 2025), UNO (Wu et al., 2025c), XVerse (Chen et al., 2025a), and

USO (Wu et al., 2025b) further improve image customization through unified multimodal modeling. MOSAIC (She et al., 2025) learns correspondence-aware alignment and multi-reference disentanglement, while UniVerse (Phung et al., 2026) studies segmentation-free multi-concept personalization. These approaches eliminate subject-specific optimization but remain highly dependent on large-scale subject-image training data, where fine-grained identity supervision is often limited.

2.3 Post-training Subject Consistency Refinement

Recently, post-training refinement has emerged as a promising paradigm for further improving existing subject-to-image models. FocusDPO (Jin et al., 2026) introduces dynamic preference optimization to enhance subject fidelity and suppress attribute leakage in pretrained personalized generation models. PSR (Wang et al., 2026) further improves multi-subject personalization by introducing pairwise subject-consistency rewards together with reinforcement learning for post-training optimization. UMO (Cheng et al., 2026) scales post-training optimization through a multi-to-multi matching paradigm with large-scale synthetic and real multi-identity customization data. USO (Wu et al., 2025b) extends this direction by incorporating Style Reward Learning into a unified framework for subject- and style-driven generation.

Different from these approaches, which rely on large-scale preference optimization or reward learning, CopyCat performs a one-time refinement completed within only a few seconds using an exact self-reconstruction objective constructed from a single proxy image. By introducing a lightweight Fine-grained Consistency LoRA, our method effectively improves fine-grained subject consistency without requiring additional large-scale training data. The resulting refined model can be directly applied to unseen reference subjects without further subject-specific optimization.

3 Method

Figure 2 illustrates the proposed framework, which consists of two independent components. First, we introduce a lightweight Fine-grained Consistency LoRA (FCLoRA), which performs a one-time refinement of a pretrained subject-to-image model within only a few seconds to improve fine-grained subject consistency. Second, we revisit the design of Subject-to-Image LoRA and propose a Visual-only LoRA strategy. Specifically, we remove LoRA from the text stream of the Double-Stream Blocks while retaining it in the image stream and the subsequent Single-Stream Blocks.

3.1 Fine-grained Consistency LoRA

3.1.1 Motivation.

Existing subject-to-image models reconstruct a target image from one or more reference images. Although the reference and target images share the same identity, they usually differ in pose, viewpoint, lighting, and many other appearance factors. Consequently, the training objective primarily encourages semantic identity alignment instead of establishing precise fine-grained correspondence between the reference and target images. As a result, current subject-to-image models generally preserve the overall identity of a subject while often failing to faithfully reproduce subtle subject-specific details. These missing details accumulate into noticeable identity inconsistency during generation.

Instead of improving supervision through additional large-scale synthesized data, we ask whether a pretrained subject-to-image model can recover these missing fine-grained details through an extremely lightweight refinement. Our key observation is that if the conditioning image and the reconstruction target are exactly the same image, all appearance variations disappear and the optimization objective becomes a pure fine-grained reconstruction problem. Since semantic identity alignment has already been learned during subject-to-image training, the refinement only needs to strengthen the model's ability to preserve fine-grained visual correspondence, making a lightweight refinement completed within only a few seconds sufficient. Motivated by this observation, we introduce a Fine-grained Consistency LoRA (FCLoRA), which performs a second-stage lightweight refinement on top of an existing Subject-to-Image LoRA.

3.1.2 FINE-GRAINED CONSISTENCY REFINEMENT.

Given a pretrained subject-to-image model with parameters $\theta = \theta_{\text{base}} + \Delta\theta_{\text{S2I}}$, where θ_{base} denotes the frozen text-to-image backbone and $\Delta\theta_{\text{S2I}}$ represents the pretrained Subject-to-Image LoRA, we further attach a lightweight Fine-grained Consistency LoRA $\Delta\theta_{\text{FC}}$, whose rank is significantly smaller than that of the original Subject-to-Image LoRA.

Unlike conventional subject-to-image training, where the conditioning image and reconstruction target depict the same subject but differ in appearance, our refinement adopts an exact self-reconstruction objective by constructing $I_{\text{ref}} = I_{\text{target}}$, where the conditioning image and the reconstruction target are exactly identical, including both pixel values and image resolution. The refinement objective is formulated as

$$\mathcal{L}_{FC} = \mathcal{L}_{FM}\left(f_{\theta_{\text{base}}+\Delta\theta_{\text{S2I}}+\Delta\theta_{\text{FC}}}(I_{\text{ref}}), I_{\text{target}}\right), \tag{1}$$

where $\mathcal{L}_{FM}$ is the original Flow Matching objective.

Compared with conventional subject-to-image training, the proposed objective completely removes appearance discrepancies between the conditioning image and the reconstruction target. Consequently, the refinement no longer needs to model pose variation, viewpoint transformation, illumination changes, or other semantic correspondence, allowing the lightweight FCLoRA to focus exclusively on strengthening fine-grained visual correspondence.

In the comparison in Figure 6(a), using the identical image as the reconstruction target improves subject consistency, whereas using another image of the same subject brings no improvement. Owing to the greatly simplified objective, the proposed FCLoRA completes the refinement within only a few seconds while consistently improving fine-grained subject consistency.

3.1.3 GENERALIZATION.

Although FCLoRA is optimized using an exact self-reconstruction proxy task, the benefit of refinement extends beyond the identity and appearance of the proxy subject. The proxy image provides a stable optimization objective that exposes fine-grained consistency errors in the pretrained subject-to-image model. We interpret the lightweight model-level refinement as correcting these errors and strengthening the model's ability to preserve subtle visual details during subject-conditioned generation. The refined model improves subject consistency for unseen identities and prompts without requiring subject-specific optimization.

3.2 VISUAL-ONLY SUBJECT-TO-IMAGE LORA

Existing subject-to-image models built upon the FLUX (Black Forest Labs, 2024) backbone typically apply LoRA to both the text and image streams of the Double-Stream Blocks, as well as the subsequent Single-Stream Blocks. However, subject-specific knowledge is primarily encoded in visual appearance, while the textual prompt mainly specifies the desired scene or attributes. This motivates us to revisit whether adapting the text stream is actually necessary for subject-to-image learning. Formally, the original Subject-to-Image LoRA is

$$\Delta\theta_{\text{S2I}} = \Delta\theta_{\text{D}}^{\text{text}} + \Delta\theta_{\text{D}}^{\text{image}} + \Delta\theta_{\text{S}}, \tag{2}$$

where $\Delta\theta_{\text{D}}^{\text{text}}$ and $\Delta\theta_{\text{D}}^{\text{image}}$ denote the LoRA parameters inserted into the text and image streams of the Double-Stream Blocks, respectively, while $\Delta\theta_{\text{S}}$ denotes the LoRA parameters in the subsequent Single-Stream Blocks.

Instead of adapting both streams, we remove LoRA only from the text stream while retaining all visual adaptation modules, resulting in

$$\Delta\theta_{\text{S2I}}^{\text{VO}} = \Delta\theta_{\text{D}}^{\text{image}} + \Delta\theta_{\text{S}}. \tag{3}$$

All remaining training settings are kept unchanged.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

4.1.1 BENCHMARKS.

We evaluate on DreamBench (Ruiz et al., 2023) and XVerseBench (Chen et al., 2025a). On DreamBench, we report I_{DINO} and I_{CLIP} to measure subject consistency, and T_{CLIP} to measure image–text alignment. XVerseBench uses 90 single-subject and 210 multi-subject prompts, with four images per prompt. We report ID for human identity and IP for general subject consistency on both splits.

4.1.2 EXPERIMENTAL PROTOCOL.

Tables 1–3 evaluate official UNO, UNO-based UMO (an RL-trained LoRA), USO, and MOSAIC subject-512 checkpoints. UMO is labeled UNO+UMO in Figure 4. Refining each model requires only one proxy image and an empty prompt. The refined model is then reused across evaluation subjects. The four-step rank-4 refinement takes approximately 3 seconds of training time, excluding model loading. Supplementary Sections A and D provide configurations and proxy comparisons. Table 4 separately studies models retrained for 20K steps on four GPUs.

Table 1: **Comparison on XVerseBench.** We evaluate MOSAIC using the officially released 512-resolution checkpoint. Its paper (She et al., 2025) reports results from a 1024-resolution model, whose checkpoint has not been released.

Method	ID_{single} ↑		IP_{single} ↑		ID_{multi} ↑		IP_{multi} ↑	
MIP-Adapter†	39.59		71.97		24.58		57.00	
OmniGen†	76.51		78.46		55.53		62.32	
OmniGen2†	62.41		74.08		40.81		67.15	
DreamO†	75.48		70.84		50.24		64.63	
XVerse†	79.48		76.86		66.59		71.48	
UniVerse‡	82.77		75.88		**66.69**		71.60	
UNO	59.68		80.46		37.53		68.05	
UNO+CopyCat	70.45	(+10.77)	**85.83**	(+5.37)	45.08	(+7.56)	**73.30**	(+5.25)
MOSAIC	59.54		74.59		36.11		66.83	
MOSAIC+CopyCat	80.97	(+21.43)	82.77	(+8.19)	48.11	(+12.00)	71.83	(+5.00)
USO	60.86		79.31		43.11		62.86	
USO+CopyCat	75.85	(+14.99)	80.21	(+0.90)	47.99	(+4.89)	65.27	(+2.41)
UMO	76.22		78.00		58.87		64.84	
UMO+CopyCat	**85.01**	(+8.78)	84.20	(+6.20)	64.49	(+5.62)	70.78	(+5.94)

†: Results from XVerse (Chen et al., 2025a); ‡: results from UniVerse (Phung et al., 2026). Other rows are evaluated by us.

4.2 COMPARISONS

4.2.1 SUBJECT-DRIVEN GENERATION.

CopyCat is designed as a lightweight refinement strategy for existing subject-to-image models. We evaluate it on four representative personalization models, including UNO, MOSAIC, USO, and UMO. Figure 3 and Tables 1, 2, and 3 present the qualitative and quantitative comparisons on both single-subject and multi-subject generation. Compared with their corresponding backbone models, CopyCat consistently preserves finer appearance details while producing more faithful multi-subject compositions. Consistent improvements in subject consistency are observed across all four backbones on both XVerseBench and DreamBench with only approximately three seconds of refinement training, demonstrating the effectiveness and efficiency of the proposed Fine-grained Consistency LoRA across different subject-to-image models.

Table 2: **Single-subject comparison on DreamBench.**

Method	I_{DINO} ↑	I_{CLIP} ↑	T_{CLIP} ↑
RealCustom++†	0.702	0.794	0.318
OmniGen†	0.693	0.801	0.315
OminiControl†	0.684	0.799	0.312
UNO	0.754	0.837	0.303
UNO+CopyCat	0.839	0.876	0.294
MOSAIC	0.644	0.795	0.318
MOSAIC+CopyCat	0.799	0.859	0.298
USO	0.750	0.828	**0.320**
USO+CopyCat	0.806	0.850	0.314
UMO	0.754	0.835	0.298
UMO+CopyCat	**0.865**	**0.893**	0.286

†: Results from UNO (Wu et al., 2025c); other rows are evaluated by us.

Table 3: **Multi-subject comparison on DreamBench.**

Method	I_{DINO} ↑	I_{CLIP} ↑	T_{CLIP} ↑
Subject Diffusion†	0.506	0.696	0.310
MIP-Adapter†	0.482	0.726	0.311
MS-Diffusion†	0.525	0.726	0.319
OmniGen†	0.511	0.722	0.331
UNO	0.531	0.732	0.320
UNO+CopyCat	0.569	0.746	0.313
MOSAIC	0.494	0.724	0.329
MOSAIC+CopyCat	0.536	0.749	0.303
USO	0.496	0.719	**0.340**
USO+CopyCat	0.521	0.728	0.332
UMO	0.527	0.728	0.316
UMO+CopyCat	**0.572**	**0.750**	0.308

†: Results from UNO (Wu et al., 2025c); other rows are evaluated by us.

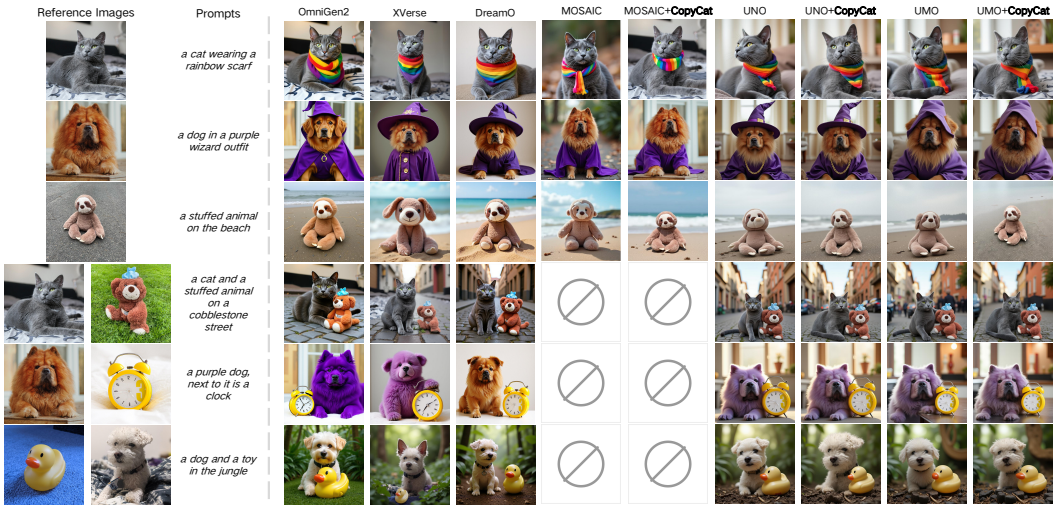


Figure 3: **Qualitative comparisons on single-subject and multi-subject personalization.** Compared with representative subject-to-image personalization models, CopyCat better preserves fine-grained subject appearance while faithfully following the target prompt. MOSAIC and MOSAIC+CopyCat are shown only for single-subject examples. Their multi-subject qualitative results are omitted because we obtained unpromising results using the officially available 512-resolution checkpoint. Zoom in for better visualization.

Beyond consistently improving existing models, the refined models also achieve strong subject-consistency performance compared with representative state-of-the-art subject-to-image methods. In particular, CopyCat achieves the highest scores on most of the reported subject-consistency metrics, demonstrating the effectiveness of the proposed refinement strategy.

4.2.2 Human Face Personalization.

Although the additional FCLoRA refinement uses no face-specific supervision, we further evaluate its performance on human face personalization. As reported in Table 1, CopyCat consistently improves the identity similarity metric (ID) over baselines. Figure 4 shows examples in which UNO+CopyCat achieves facial identity preservation visually comparable to the UNO-based UMO



Figure 4: **Comparison of Human Identity Consistency Improvement.** UNO+CopyCat shows facial identity preservation visually comparable to UMO, without face-specific supervision during FCLoRA refinement.



Figure 5: **Alleviating identity confusion between similar-category subjects.** CopyCat mitigates identity confusion when composing visually similar subjects by preserving the fine-grained characteristics of each subject.

model. These results demonstrate that the proposed refinement strategy naturally generalizes to human face personalization.

4.3 Further Analysis

4.3.1 Identity Confusion Mitigation.

Identity confusion is a common failure mode in multi-subject personalization, particularly for visually similar subjects. As shown in Figure 5, UNO often mixes the identities of different subjects, leading to ambiguous appearances. CopyCat effectively separates individual identities and preserves their distinctive visual characteristics, substantially reducing identity confusion.

4.3.2 Effect of the Refinement Target.

Figure 6 analyzes the refinement target and duration. As shown in Figure 6(a), using another image of the same subject as the reconstruction target brings no improvement, whereas the proposed identical-image target noticeably improves subject consistency within only a few refinement steps. Longer refinement gradually overfits to image-specific details, suggesting that only a lightweight refinement is necessary. Figure 6(b) further demonstrates that the refinement benefits an unseen subject beyond the proxy subject.

4.3.3 Effect of Visual-only LoRA and FCLoRA.

Table 4 compares the conventional Full LoRA with the proposed Visual-only LoRA. Since both models are retrained on our available training data, their absolute performance is not directly comparable to the official UNO checkpoint. Nevertheless, Visual-only LoRA consistently improves subject consistency on both benchmarks. Additional qualitative comparisons are provided in the supplementary material. These results support removing LoRA from the dedicated text stream while retaining adaptation in the image branch and the subsequent shared Single-Stream Blocks in this setting.

We further apply FCLoRA to the Visual-only model. As shown in Table 4, FCLoRA further improves human identity and general subject consistency on XVerseBench, as well as DINO and CLIP image similarity on DreamBench. These results show that the proposed refinement remains effective when combined with Visual-only LoRA.

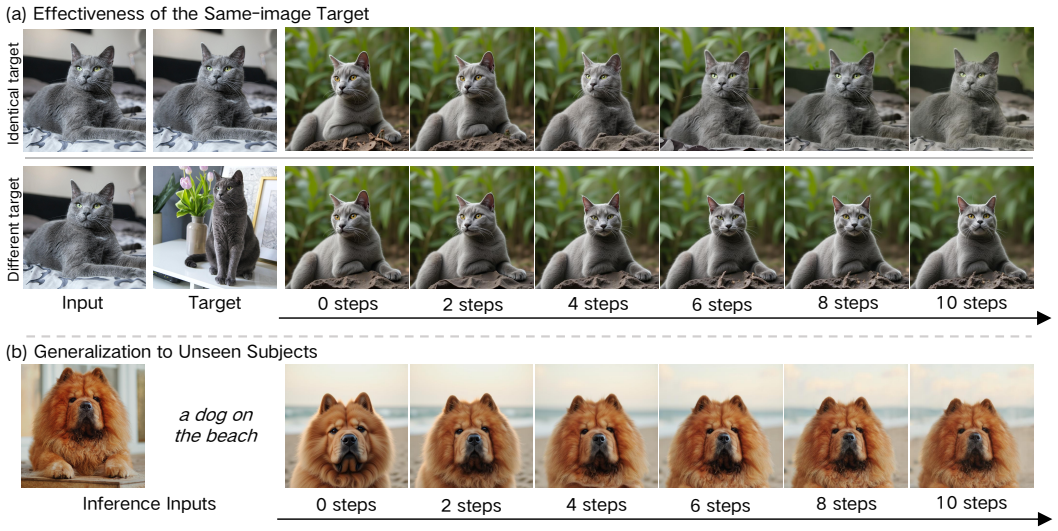


Figure 6: **Importance of the same-image refinement target.** (a) Top: identical-image target; bottom: a different image of the same subject. Identical-image refinement improves fine-grained consistency within a few steps; the different-image target brings no improvement. Excessive optimization can overfit image-specific appearance. (b) Generation on an unseen subject using the models obtained at the corresponding training steps from the identical-target setting in the top row of (a), without further optimization on the new subject. All step counts denote training steps.

Table 4: **Effect of Visual-only LoRA and FCLoRA.**

		XVerseBench		DreamBench	
Visual-only	FCLoRA	ID_{single} ↑	IP_{single} ↑	I_{DINO} ↑	I_{CLIP} ↑
—	—	29.13	70.17	0.6562	0.7908
✓	—	39.30	73.14	0.7033	0.8157
✓	✓	**54.00**	**77.32**	**0.8271**	**0.8794**

5 Conclusion

In this paper, we presented CopyCat, a lightweight refinement framework for improving fine-grained subject consistency in subject-to-image personalization models. We showed that a simple proxy task based on identical image reconstruction is sufficient to learn a transferable fine-grained consistency capability. This capability consistently enhances identity preservation across multiple personalization models after only a few seconds of refinement. We further demonstrated that a simple Visual-only LoRA design consistently improves subject consistency by removing dedicated text-stream LoRA while retaining image-stream and shared-block adaptation. Experiments on XVerseBench and DreamBench validate FCLoRA across the evaluated subject-to-image models and the effectiveness of the Visual-only design. We hope the proposed refinement strategy provides a simple and practical direction for improving future subject-to-image personalization models.

AI use statement

We used GPT to polish the writing of this paper. The authors reviewed and verified the AI-polished text and take full responsibility for the final content of this paper.

References

Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–12, 2023.

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

Bowen Chen, Mengyi Zhao, Haomiao Sun, Li Chen, Xu Wang, Kang Du, and Xinglong Wu. XVerse: Consistent multi-subject control of identity and semantic attributes via DiT modulation. *arXiv preprint arXiv:2506.21416*, 2025a.

Li Chen, Mengyi Zhao, Yiheng Liu, Mingxu Ding, Yangyang Song, Shizun Wang, Xu Wang, Hao Yang, Jing Liu, Kang Du, et al. Photoverse: Tuning-free image customization with text-to-image diffusion models. *arXiv preprint arXiv:2309.05793*, 2023a.

Nan Chen, Mengqi Huang, Zhuowei Chen, Yang Zheng, Lei Zhang, and Zhendong Mao. Customcontrast: A multilevel contrastive perspective for subject-driven text-to-image customization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2123–2131, 2025b. doi: 10.1609/aaai.v39i2.32210.

Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36:30286–30305, 2023b.

Yufeng Cheng, Wenxu Wu, Shaojin Wu, Mengqi Huang, Fei Ding, and Qian He. Scaling multi-identity consistency for image customization via multi-to-multi matching paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1906–1916, 2026.

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

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. *Advances in Neural Information Processing Systems*, 36:15890–15902, 2023.

Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, Peng Zhang, and Qian He. Pulid: Pure and lightning id customization via contrastive alignment. *Advances in neural information processing systems*, 37:36777–36804, 2024.

Junjie He, Yuxiang Tuo, Binghui Chen, Chongyang Zhong, Yifeng Geng, and Liefeng Bo. Anystory: Towards unified single and multiple subject personalization in text-to-image generation. *arXiv preprint arXiv:2501.09503*, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Qihan Huang, Siming Fu, Jinlong Liu, Hao Jiang, Yipeng Yu, and Jie Song. Resolving multi-condition confusion for finetuning-free personalized image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3707–3714, 2025.

Xuhui Jia, Yang Zhao, Kelvin CK Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*, 2023.

Qiaoqiao Jin, Siming Fu, Dong She, Weinan Jia, Hualiang Wang, Mu Liu, and Jidong Jiang. Focusdpo: Dynamic preference optimization for multi-subject personalized image generation via adaptive focus. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5512–5520, 2026.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1931–1941, 2023.

Gihyun Kwon, Simon Jenni, Dingzeyu Li, Joon-Young Lee, Jong Chul Ye, and Fabian Caba Heilbron. Concept weaver: Enabling multi-concept fusion in text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8880–8889, 2024.

Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36:30146–30166, 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8640–8650, 2024.

Habin Lim, Yeongseob Won, Juwon Seo, and Gyeong-Moon Park. Conceptsplit: Decoupled multi-concept personalization of diffusion models via token-wise adaptation and attention disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18421–18430, 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 *Proceedings of the 40th International Conference on Machine Learning*, pp. 21548–21566, 2023a.

Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Customizable image synthesis with multiple subjects. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 57500–57519. Curran Associates, Inc., 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/b3847cda0c8cc0cfcdacf462dc122214-Paper-Conference.pdf.

Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024.

Zhendong Mao, Mengqi Huang, Fei Ding, Mingcong Liu, Qian He, and Yongdong Zhang. Realcustom++: Representing images as real textual word for real-time customization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Tuna Han Salih Meral, Enis Simsar, Federico Tombari, and Pinar Yanardag. Contrastive test-time composition of multiple lora models for image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18090–18100, 2025.

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.

Quynh Phung, Sandesh Ghimire, Minsi Hu, Chung-Chi Tsai, and Jia-Bin Huang. UniVerse: A unified modulation framework for segmentation-free, disentangled multi-concept personalization. *arXiv preprint arXiv:2606.00351*, 2026.

Ryan Po, Guandao Yang, Kfir Aberman, and Gordon Wetzstein. Orthogonal adaptation for modular customization of diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7964–7973, 2024.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–22510, 2023.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6527–6536, 2024.

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 *European Conference on Computer Vision*, pp. 422–438. Springer, 2024.

Dong She, Siming Fu, Mushui Liu, Qiaoqiao Jin, Hualiang Wang, Mu Liu, and Jidong Jiang. MOSAIC: Multi-subject personalized generation via correspondence-aware alignment and disentanglement. *arXiv preprint arXiv:2509.01977*, 2025.

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8543–8552, 2024.

Yair Shpitzer, Gal Chechik, and Idan Schwartz. Single image iterative subject-driven generation and editing. *arXiv preprint arXiv:2503.16025*, 2025.

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

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

Kuan-Chieh Wang, Daniil Ostashev, Yuwei Fang, Sergey Tulyakov, and Kfir Aberman. Moa: Mixture-of-attention for subject-context disentanglement in personalized image generation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–12, 2024a.

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*, 2024b.

Shulei Wang, Longhui Wei, Xin He, Jianbo Ouyang, Hui Lu, Zhou Zhao, and Qi Tian. Psr: Scaling multi-subject personalized image generation with pairwise subject-consistency rewards. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14756–14766, June 2026.

Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. In *International Conference on Learning Representations*, volume 2025, pp. 95118–95146, 2025.

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15943–15953, 2023.

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*, 2025a.

Shaojin Wu, Mengqi Huang, Yufeng Cheng, Wenxu Wu, Jiahe Tian, Yiming Luo, Fei Ding, and Qian He. Uso: Unified style and subject-driven generation via disentangled and reward learning. *arXiv preprint arXiv:2508.18966*, 2025b.

Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18682–18692, 2025c.

Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, 133(3):1175–1194, 2025a.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13294–13304, 2025b.

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.

Peng Zheng, Ye Wang, Rui Ma, and Zuxuan Wu. Freelora: Enabling training-free lora fusion for autoregressive multi-subject personalization. *arXiv preprint arXiv:2507.01792*, 2025.