Personalize Anything for Free with Diffusion Transformer

Haoran Feng[1*], **Zehuan Huang**[2*], **Lin Li**[3], **Lu Sheng**[2†]

[1]Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2]School of Software, Beihang University, Beijing, China
[3]School of Finance, Renmin University, Beijing, China
fenghr24@mails.tsinghua.edu.cn, huangzehuan@buaa.edu.cn, 2019200839@ruc.edu.cn, lsheng@buaa.edu.cn

Abstract

Personalized image generation aims to produce images of user-specified concepts while enabling flexible editing. Recent training-free approaches, while exhibiting higher computational efficiency than training-based methods, struggle with identity preservation, applicability, and compatibility with diffusion transformers (DiTs). In this paper, we uncover the untapped potential of DiT, where simply replacing denoising tokens with those of a reference subject achieves zero-shot subject reconstruction. This simple yet effective feature injection technique unlocks diverse scenarios, from personalization to image editing. Building upon this observation, we propose *Personalize Anything*, a training-free framework that achieves personalized image generation in DiT through: 1) timestep-adaptive token replacement that enforces subject consistency via early-stage injection and enhances flexibility through late-stage regularization, and 2) patch perturbation strategies to boost structural diversity. Our method seamlessly supports layout-guided generation, multi-subject personalization, and mask-controlled editing. Evaluations demonstrate that our method, without requiring any training, achieves state-of-the-art performance in identity preservation and versatility. Our work establishes new insights into DiTs while delivering a practical paradigm for efficient personalization.

Code — https://github.com/fenghora/personalize-anything

1 Introduction

Personalized image generation aims to synthesize images of user-specified concepts while enabling flexible editing. The advent of text-to-image diffusion models (Ramesh et al. 2022; Saharia et al. 2022; Podell et al. 2023; Labs 2024; Rombach et al. 2022) has revolutionized this field, enabling applications in areas like advertising production.

Previous research on subject image personalization relies on test-time optimization or large-scale fine-tuning. The optimization-based approach (Ruiz et al. 2023; Kumari et al. 2023; Gu et al. 2024) fine-tunes pre-trained models on a few subject images to learn the specific concept.

While achieving identity preservation, these methods demand substantial computational resources and time due to per-subject optimization over hundreds of iterations. Large-scale fine-tuning alternatives like (Li, Li, and Hoi 2024; Ye et al. 2023; Ma et al. 2024) aim to address this by training auxiliary networks on large datasets to encode reference images. However, these approaches both demand heavy training requirements and risk overfitting to narrow data distributions, degrading their generalizability.

Figure 1: Simple token replacement or resizing in DiT (right) achieves high-fidelity subject reconstruction via its position-disentangled representation, while U-Net's convolutional entanglement (left) induces blurring and artifacts.

Recent training-free solutions (Tewel et al. 2024b; Zhou et al. 2024; Ding et al. 2024) exhibit higher computational efficiency than training-based approaches. They typically leverage attention sharing mechanisms to inject reference features, processing denoising and reference subject tokens jointly in pre-trained self-attention layers. However, these attention-based methods exhibit several limitations: 1) They often fail to preserve reference identity due to the lack of explicit constraints. 2) These methods are task-specific, limiting their versatility and hindering their application in scenarios like layout-guided generation, multi-subject personalization, and mask-controlled editing. 3) Their integration with advanced text-to-image diffusion transformers (DiTs) (Peebles and Xie 2023; Labs 2024; Esser et al. 2024) is hindered by the strong influence of encoded positional information on DiT's attention mechanism, as analyzed in Sec. 3.2.

In this paper, we delve into the diffusion transformers (DiTs) (Peebles and Xie 2023; Labs 2024; Esser et al. 2024),

*These authors contributed equally.
†Corresponding author.

Figure 2: *Personalize Anything* is a **training-free** framework based on Diffusion Transformers (DiT) for personalized image generation. The framework demonstrates advanced **versatility**, excelling in *single-subject personalization, layout-guided subject personalization (top-left)*, *multi-subject, subject-scene composition (middle-left)*, as well as applications like *visual storytelling (bottom-left), inpainting (top-right) and outpainting (bottom-right)*, all without any training or fine-tuning.

and observe that simply replacing the denoising tokens with those of a reference subject allows for high-fidelity subject reconstruction. As illustrated in Fig. 1, DiT exhibits exceptional reconstruction fidelity under this manipulation, while U-Net (Ronneberger, Fischer, and Brox 2015) often induces blurred edges and artifacts. We attribute this to the separate embedding of positional information in DiT, achieved via its explicit positional encoding mechanism. This decoupling of semantic features and position enables the substitution of purely semantic tokens, avoiding positional interference. Conversely, U-Net's convolutional mechanism binds texture and spatial position together, causing positional conflicts when replacing tokens and leading to low-quality image generation. This discovery establishes token replacement as a viable pathway for zero-shot subject personalization in DiT, unlocking various scenarios ranging from personalization to inpainting and outpainting, without necessitating complicated attention engineering.

Building on this foundation, we propose '*Personalize Anything*", a training-free framework for personalized image generation and a variety of other tasks. Our approach leverages novel timestep-adaptive token replacement and patch perturbation strategies. Specifically, we inject reference subject tokens (excluding positional information) in the earlier steps of the denoising process to enforce subject consistency, while enhancing flexibility in the later steps through multi-modal attention. Furthermore, we introduce patch perturbation to the reference tokens before token replacement, locally shuffling them and applying morphological operations to the subject mask. It encourages the model to introduce more global appearance information and enhances structural and textural diversity. Additionally, our framework seamlessly supports 1) layout-guided generation through translations on replacing regions, 2) multi-subject personalization and subject-scene composition via sequential injection of reference subjects or scene, and 3) extended applications (e.g. inpainting and outpainting) via incorporating user-specified mask conditions.

Comprehensive evaluations on multiple personalization tasks demonstrate that our training-free method exhibits superior identity preservation, fidelity and versatility, outperforming existing approaches including those fine-tuned on DiTs. Our contributions are summarized as follows:

- We uncover DiT's potential for high-fidelity subject reconstruction via simple token replacement, and characterize its position-disentangled properties.
- We introduce a simple yet effective framework, denoted as "Personalize Anything", which starts with subject reconstruction and enhances the flexibility via timestep-adaptive replacement and patch perturbation.
- Experiments demonstrate that our framework successfully handles a wide range of personalization tasks without additional training, exhibiting strong consistency, fidelity, and versatility.

2 Related Work

Text-to-Image Diffusion Models. Diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2022) have transformed text-to-image generation by progressively denoising Gaussian distributions into images. Among a series of effective works (Ramesh et al. 2022; Saharia et al. 2022;

Podell et al. 2023; Peebles and Xie 2023; Labs 2024; Rombach et al. 2022), the Latent Diffusion Model (Rombach et al. 2022) which denoises in latent space with a U-Net backbone (Ronneberger, Fischer, and Brox 2015) laid the foundation for improved resolution (Podell et al. 2023). Recently, vision transformers (Dosovitskiy 2020; Peebles and Xie 2023) have replaced U-Nets, leveraging global attention mechanisms and geometrically-aware positional encodings. These diffusion transformers (DiTs) (Esser et al. 2024; Labs 2024) offer superior scalability and performance improvements, establishing them as the new state-of-the-art.

Personalized Image Generation. Previous subject personalization methods primarily adopt three strategies: i) Test-time optimization techniques (Ruiz et al. 2023; Gal et al. 2022; Kumari et al. 2023; Gu et al. 2024; Liu et al. 2023; Kwon et al. 2024; Kong et al. 2024; Zhu et al. 2024; Shah et al. 2024; Hyung, Shin, and Choo 2024; Voynov et al. 2023; Han et al. 2023; Tewel et al. 2024a; Ni et al. 2025) that fine-tune foundation models on target concepts at inference time, often requiring 30 GPU-minute optimization per subject; ii) large-scale training-based methods (Li, Li, and Hoi 2024; Ye et al. 2023; Ma et al. 2024; Zhang et al. 2024; Patel et al. 2024; Jiang et al. 2024; Song et al. 2024; Wang et al. 2025; Huang et al. 2024; Tan et al. 2024; Le et al. 2024; Chen et al. 2024b; Zong et al. 2024; He et al. 2025; Chen et al. 2024a; Wang et al. 2025; Liu et al. 2025; Cai et al. 2024; Tao et al. 2025) that learn concept embeddings through auxiliary networks pre-trained on large datasets. While achieving notable fidelity, both paradigms suffer from computational overheads and distribution shifts that limit real-world application. iii) To exhibit higher computational efficiency, training-free methods (Tewel et al. 2024b; Zhou et al. 2024; Ding et al. 2024; Shin et al. 2024) that typically leverage an attention sharing mechanism to inject reference features, processing denoising and reference subject tokens jointly in pre-trained self-attention layers. However, these attention-based methods lack subject consistency and struggle to preserve identity. Moreover, explicit positional encoding in DiTs (Peebles and Xie 2023) limits their scalability to larger text-to-image models (Labs 2024; Esser et al. 2024), and they are often restricted to a single application.

3 Methodology

This paper presents a training-free approach for personalized generation with diffusion transformers (DiTs) (Peebles and Xie 2023), enabling high-fidelity depictions of user concepts while maintaining textual controllability. We begin with an overview of standard architectures in text-to-image diffusion models; Sec. 3.2 highlights distinctions that hinder the application of attention sharing mechanisms to DiTs; Sec. 3.3 demonstrates DiT's potential for subject reconstruction via simple token replacement, leading to the introduction of our *Personalize Anything* framework in Sec. 3.4.

3.1 Preliminaries

Diffusion models progressively denoise a latent variable z_T through a network ϵ_θ, with architectural choices being of paramount importance. We analyze two main paradigms:

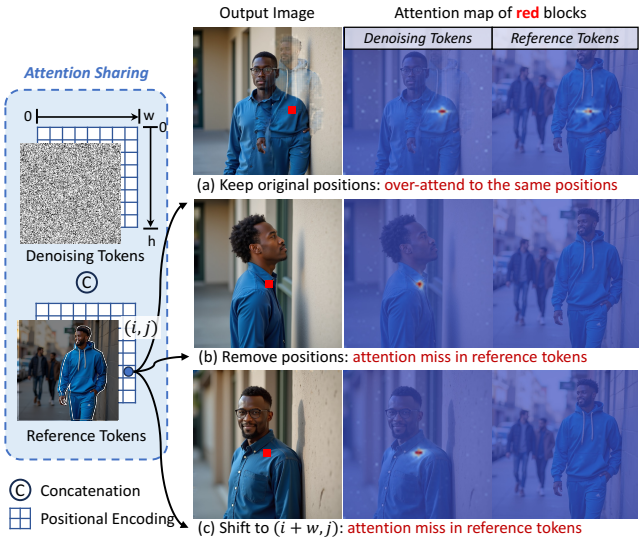

Figure 3: Attention sharing (Tewel et al. 2024b; Ding et al. 2024) fails in DiT due to the explicit positional encoding mechanism. When keeping the original positions $(i, j) \in [0, w) \times [0, h)$ in reference tokens, denoising tokens over-attend to reference ones with the same positions (shown in attention maps of (a)), resulting in ghosting artifacts in the generated image. Modified strategies, (b) removing positions and (c) shifting to non-overlapping regions, avoid collisions but loses identity alignment, as attention is almost absent on reference tokens.

U-Net (Ronneberger, Fischer, and Brox 2015) in Stable Diffusion (Rombach et al. 2022) comprises pairs of down-sampling and up-sampling blocks connected by a middle block. Each block interleaves residual blocks for feature extraction with spatial attention layers for capturing spatial relationships.

Diffusion Transformer (DiT). Modern DiTs (Peebles and Xie 2023) in advanced models (Labs 2024; Esser et al. 2024) leverage transformers (Dosovitskiy 2020) to process discretized latent representations, including image tokens $X \in \mathbb{R}^{N \times d}$ and text tokens $C \in \mathbb{R}^{M \times d}$, where d is the embedding dimension, N and M are the length of sequences. These models typically encode positional information of X through RoPE (Su et al. 2024), which applies rotation matrices based on the token's coordinate (i, j) in the 2D grid. Text tokens C receive fixed positional anchors $(i = 0, j = 0)$ to maintain modality distinction. The multi-modal attention mechanism (MMA) is then applied to all position-encoded tokens $[\dot{X}; \dot{C}_T] \in \mathbb{R}^{(N+M) \times d}$, enabling full bidirectional attention across both modalities.

3.2 Attention Sharing Fails in DiT

We systematically investigate why established U-Net-based personalization techniques (Ding et al. 2024; Tewel et al. 2024b) fail when naively applied to DiT architectures (Labs 2024), identifying positional encoding conflicts as the core challenge.

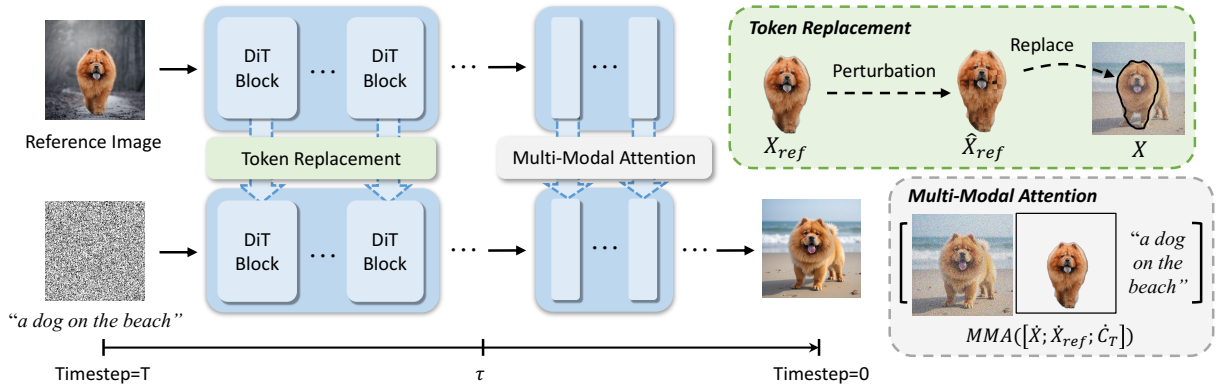

Figure 4: Method overview. Our framework anchors subject identity through mask-guided token replacement with preserved positional encoding in early denoising, then transitions to multi-modal attention for semantic fusion with text. During token replacement, we inject variations via patch perturbations, balancing identity preservation and generative flexibility.

Positional Encoding Collision. To share attention as in prior work (Tewel et al. 2024b; Ding et al. 2024) in DiT, we concatenate position-encoded denoising tokens $\dot{X}$ and reference tokens $\dot{X}_{ref}$ (obtained via flow inversion (Rout et al. 2024)) into a unified sequence $[\dot{X}; \dot{X}_{ref}]$. Both tokens keep the original positions $(i, j) \in [0, w) \times [0, h)$, causing destructive interference to attention computation. As visualized in Fig. 3a, this forces denoising tokens to over-attend to reference tokens with the same positions, resulting in ghosting artifacts of the reference subject in the generated image.

Modified Encoding Strategies. Motivated by DiT's position-disentangled encoding (Sec. 3.1), we engineer two positional adjustments on X_{ref} to obtain non-conflicting $\dot{X}_{ref}$: i) remove positions and fix all reference positions to $(0, 0)$ akin to text tokens, and ii) shift reference tokens to $(i', j') = (i + w, j)$, creating non-overlapping regions. As shown in Fig. 3 (b) and (c), while eliminating collisions, both methods struggle to preserve identity, as attention is almost absent on reference tokens.

In summary, the explicitly encoded positional information exhibits strong influence on the attention mechanism in DiT—a fundamental divergence from U-Net's implicit position handling. This makes it difficult for generated images to correctly attend to the reference subject's tokens within traditional attention sharing.

3.3 Token Replacement in DiT

Building on the foundational observation on DiT's architectural distinctions, we extend our investigation to the latent representation in DiT. We uncover that simply replacing the denoising tokens with those of a reference subject allows for high-fidelity subject reconstruction. Specifically, we apply inversion techniques (Rout et al. 2024) on the reference image, obtaining the reference tokens X_{ref} without encoded positional information, as well as the reference subject's mask $\mathcal{M}_{ref}$. We then inject X_{ref} into specific region $\mathcal{M}$ of the denoising tokens X via token replacement:

$$\hat{X} = X \odot (1 - \mathcal{M}) + X_{ref} \odot \mathcal{M} \quad (1)$$

where $\mathcal{M}$ can be obtained by translating $\mathcal{M}_{ref}$. As shown in Fig. 1, token replacement in DiT reconstructs high-fidelity images with consistent subjects in specified positions, while U-Net's convolutional entanglement manifests as blurred edges and artifacts.

We attribute this to the separate embedding of positional information in DiT, achieved via its explicit positional encoding mechanism (Sec. 3.1). This decoupling of semantic features and position enables the substitution of purely semantic tokens, avoiding positional interference. Conversely, U-Net's convolutional mechanism binds texture and spatial position together, causing positional conflicts when replacing tokens and leading to low-quality image generation. This discovery establishes token replacement as a viable pathway for zero-shot subject personalization in DiT. It unlocks various scenarios ranging from personalization to inpainting and outpainting, without necessitating complicated attention engineering, and establishes the foundation for our personalization framework in Sec. 3.4.

3.4 Personalize Anything

Building on these findings, we propose *Personalize Anything*, a novel training-free framework for flexible personalization across diverse tasks. Inspired by zero-shot subject reconstruction in DiTs, our approach employs timestep-adaptive token replacement and patch perturbation to enhance adaptability (Fig. 4).

Timestep-adaptive Token Replacement. Our method begins by inverting a reference image containing the desired subject (Rout et al. 2024). This process yields reference tokens X_{ref} (excluding positional encodings) and a corresponding subject mask $\mathcal{M}_{ref}$ (Wang et al. 2023). Instead of continuous injection throughout the denoising process as employed in subject reconstruction, we introduce a timestep-dependent strategy:

1) Early-stage subject anchoring via token replacement **($t > \tau$).** During the initial denoising steps ($t > \tau$, where τ is an empirically determined threshold set at 80% of the total denoising steps T), we anchor the subject's identity by re-

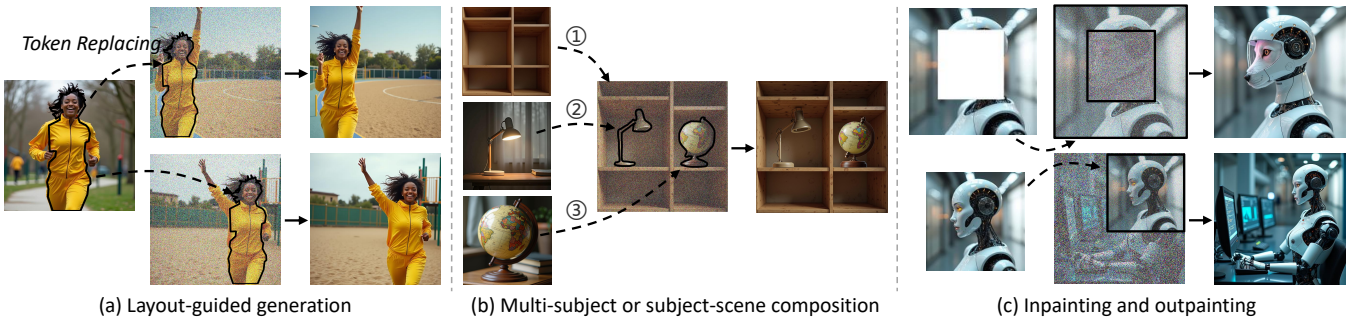

Figure 5: Seamless extensions. Our framework enables: (a) layout-guided generation by replacing injected tokens, (b) multi-subject composition via sequential token injection, and (c) inpainting and outpainting with masks and increased replacement.

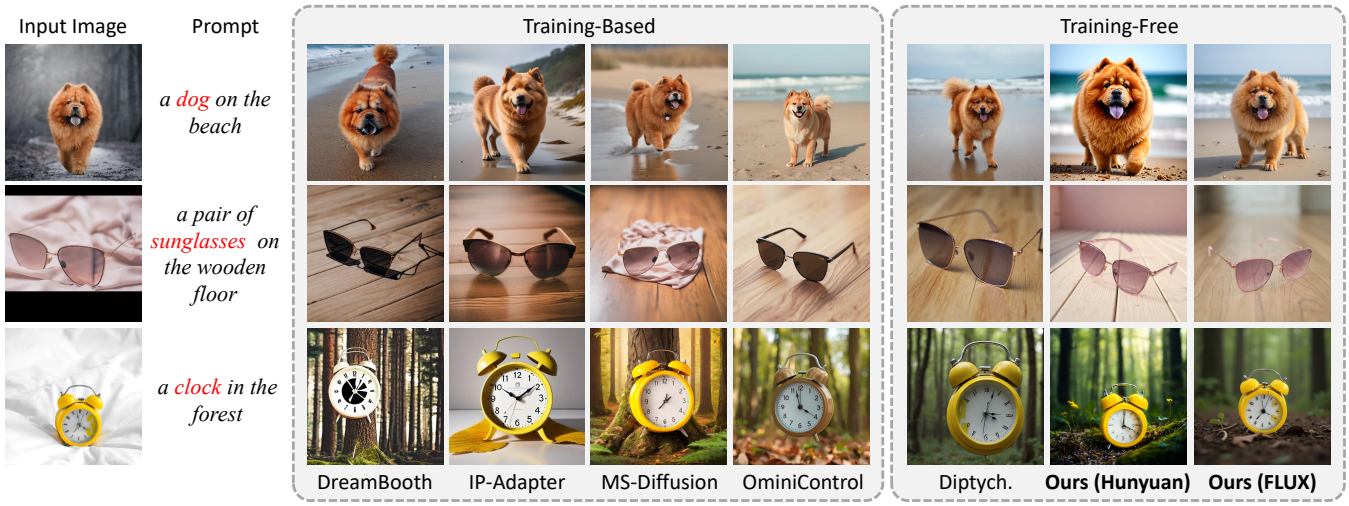

Figure 6: Qualitative comparisons on single-subject personalization.

placing the denoising tokens X within the subject region $\mathcal{M}$ with the reference tokens X_{ref} (eq. (1)). The region $\mathcal{M}$ can be obtained by translating $\mathcal{M}_{ref}$ to the user-specified location. We preserve the positional encodings associated with the denoising tokens X to maintain spatial coherence.

2) Later-stage semantic fusion via multi-modal attention ***($t \leq \tau$).*** In later denoising steps $t \leq \tau$, we switch to semantic fusion by concatenating zero-positioned reference tokens $\dot{X}ref$ with denoising tokens $\dot{X}$ and text embeddings $\dot{C}$. The unified sequence $[\dot{X}; \dot{X}_{ref}; \dot{C}_T]$ then undergoes Multi-Modal Attention (MMA) to harmonize subject guidance with textual conditions. This adaptive threshold τ balances the preservation of subject identity with the flexibility afforded by the text prompt.

Patch Perturbation for Variation. To prevent identity overfitting while preserving identity consistency, we introduce two complementary perturbations: 1) Random local token shuffling within 3x3 windows disrupts rigid texture alignment, and 2) Mask augmentation of $\mathcal{M}_{ref}$, including simulating natural shape variations via morphological dilation/erosion with a 5px kernel, or manually selecting regions emphasizing identity. This local interference technique is to encourage the model to introduce more global textural features while enhancing structural and local diversity.

Seamless Extensions. As illustrated in Fig. 5, our framework naturally extends to complex scenarios without additional training. Translating $\mathcal{M}$ enables the spatial arrangement of subjects thereby achieving layout-guided generation, while sequential injection of multiple $\{X^k_{ref}\}$ into disjoint $\{\mathcal{M}_k\}$ regions and unified Multi-Modal Attention MMA$([\dot{X}; \{\dot{X}^k_{ref}\}; \dot{C}_T])$ facilitate multi-subject or subject-scene composition. Additionally, the framework also supports resizing X_{ref} to control object scale, enabling fine-grained manipulation of diverse attributes. For image editing tasks, we incorporate user-specified masks in the inversion process to obtain reference X_{ref} and $\mathcal{M}_{ref}$ that should be preserved. Our framework ensures high-fidelity reconstruction of the preserved regions, while achieving semantic alignment between the edited areas and the input text.

4 Experiments

4.1 Experimental Setup

Implementation Details. Our framework builds upon HunyuanDiT (Li et al. 2024) and FLUX.1-dev (Labs 2024).

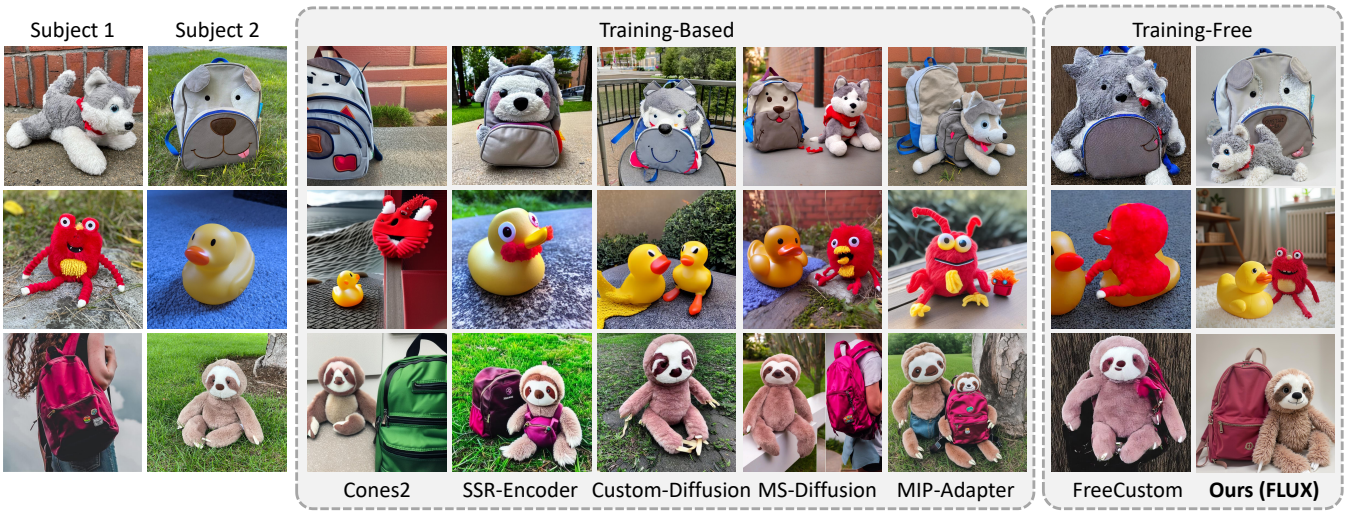

Figure 7: Qualitative comparisons on multi-subject personalization.

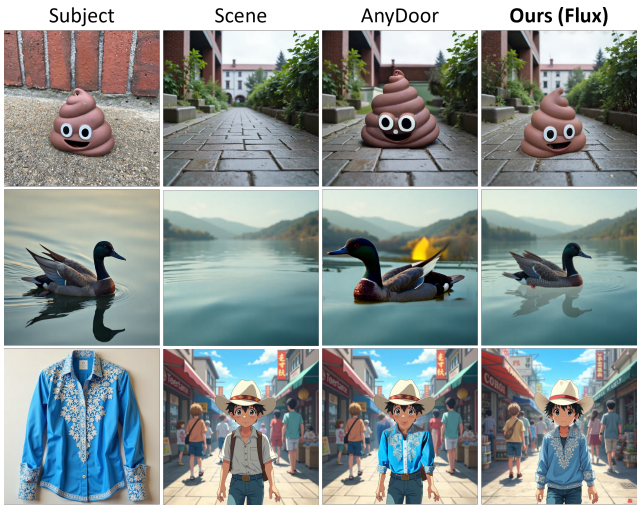

Figure 8: Qualitative results on subject-scene composition.

We adopt 50-step sampling with classifier-free guidance ($w = 3.5$), generating 1024×1024 resolution images. Token replacement threshold τ is set to 80% total steps.

Benchmark Protocols. We establish three evaluation tiers: 1) Single-subject personalization, compared against 10 approaches spanning training-based (Wang et al. 2025; Tan et al. 2024; Le et al. 2024; Duan et al. 2024; Patel et al. 2024; Zhang et al. 2024; Li, Li, and Hoi 2024; Ye et al. 2023; Ruiz et al. 2023) and training-free (Tewel et al. 2024b; Shin et al. 2024) paradigms, 2) Multi-subject personalization, evaluated against 6 representative methods (Wang et al. 2025; Patel et al. 2024; Kumari et al. 2023; Liu et al. 2023; Huang et al. 2024; Ding et al. 2024), and 3) Subject-scene composition, benchmarked using AnyDoor (Chen et al. 2024a) as reference for contextual adaptation.

Evaluation Metrics. We evaluate our *Personalize Anything* on DreamBench (Ruiz et al. 2023) which comprises 30 base objects each accompanied by 25 textual prompts. We extend this dataset to 750, 1000, and 100 test cases for single-subject, multi-subject, and subject-scene personalization using combinatorial rules. Quantitative assessment uses multi-dimensional metrics: CLIP-T (Radford et al. 2021) for image-text alignment, and DINO (Oquab et al. 2024), CLIP-I (Radford et al. 2021), DreamSim (Fu et al. 2023) for identity preservation in single-subject evaluation while SegCLIP-I (Zhu et al. 2024) in multi-subject evaluation.

4.2 Comparison to State-of-the-Arts

Single-Subject Personalization. Fig. 6 presents a qualitative comparison with representative baseline methods. Existing test-time fine-tuning methods (Ruiz et al. 2023) require optimization for each concept and may confuse the background with the subject. Training-based, test-time tuning-free methods (Ye et al. 2023; Le et al. 2024; Tan et al. 2024; Wang et al. 2025), while trained on large datasets, struggle with identity preservation in real images. Training-free methods (Tewel et al. 2024b; Shin et al. 2024) generate inconsistent subjects with single-image inputs. In contrast, our method generates high-fidelity, subject-consistent images without the need for training or fine-tuning. Quantitative results in Sec. 4.2 confirm our superior performance in identity preservation and image-text alignment.

Multi-Subject Personalization. As shown in Fig. 7, existing methods (Liu et al. 2023; Zhang et al. 2024; Kumari et al. 2023; Wang et al. 2025; Huang et al. 2024) often struggle with conceptual fusion, failing to preserve subject identities or producing fragmented results due to poor inter-subject relationship modeling. In contrast, our method maintains natural subject interactions via layout-guided generation while ensuring each subject retains its distinct identities. Quantitative results in Sec. 4.2 demonstrate the superiority of *Personalize Anything* in SegCLIP-I and CLIP-T,

Method	CLIP-T↑	CLIP-I↑	DINO↑	DreamSim↓
DreamBooth (DiT)	0.271	0.819	0.550	0.290
BLIP-Diffusion	0.251	0.835	0.641	0.283
IP-Adapter (DiT)	0.249	0.861	0.652	0.256
λ-ECLIPSE	0.235	0.866	0.682	0.224
SSR-Encoder	0.244	0.860	**0.701**	0.220
EZIGen	0.263	0.825	0.662	0.247
MS-Diffusion	0.283	0.824	0.539	0.261
OneDiffusion (DiT)	0.255	0.817	0.603	0.298
OminiControl (DiT)	0.275	0.820	0.516	0.301
ConsiStory	0.284	0.753	0.472	0.434
Diptych. (DiT)	0.301	0.863	0.502	0.231
Ours (HunyuanDiT)	0.291	0.869	0.679	0.206
Ours (FLUX)	**0.307**	**0.876**	0.683	**0.179**

Table 1: Quantitative results on single-subject personalization.

Figure 9: Qualitative ablation studies on token replacement threshold τ and patch perturbation.

effectively capturing identities and preserving text control.

Subject-Scene Composition. We evaluate *Personalize Anything* on subject-scene composition, comparing it with Anydoor (Chen et al. 2024a). As shown in Fig. 8, Anydoor generates incoherent results, with inconsistencies between subjects and factors like lighting. In contrast, our method produces natural images while preserving subject details, showcasing the potential and generalization of *Personalize Anything* for high-fidelity personalized image generation.

4.3 Ablation Study

We conduct ablation studies on single-subject personalization, examining the effects of token replacement timestep threshold τ and the patch perturbation strategy.

Effects of Threshold τ. Our study shows that the timestep threshold τ balances subject consistency and flexibility (Fig. 9). Early-to-mid replacements ($\tau > 0.8\ T$) integrate geometric and appearance priors, evolving from coarse layouts ($0.9\ T$) to refined textures ($0.8\ T$), while lower τ ($0.7\ T$) produces identical subjects. Quantitative results (Sec. 4.2) indicate $\tau = 0.8\ T$ achieves optimal reference similarity (0.882, CLIP-I) and image-text alignment (0.302, CLIP-T).

Effects of Patch Perturbation. By combining local token shuffling and mask morphing, our perturbation strategy reduces both texture and structure overfitting. With $\tau = 0.8\ T$, the generated subject and the reference one are structurally similar without perturbation (Fig. 9). Conversely, applying perturbation makes the structure and texture more flexible while maintaining identically consistency.

Method	CLIP-T↑	CLIP-I↑	SegCLIP-I↑
λ-ECLIPSE	0.258	0.738	0.757
SSR-Encoder	0.234	0.720	0.761
Cones2	0.255	0.747	0.702
Custom Diffusion	0.228	0.727	0.781
MIP-Adapter	0.276	0.765	0.751
MS-Diffusion	0.278	0.780	0.748
FreeCustom	0.248	0.749	0.734
Ours (HunyuanDiT)	0.284	0.817	0.832
Ours (FLUX)	**0.302**	**0.843**	**0.891**

Table 2: Quantitative results on multi-subject personalization.

τ	Pertur.	CLIP-T↑	CLIP-I↑	DINO↑	DreamSim↓
T	✗	**0.317**	0.764	0.625	0.305
$0.95\ T$	✗	0.313	0.773	0.632	0.294
$0.90\ T$	✗	0.306	0.849	0.680	0.199
$0.80\ T$	✗	0.302	0.882	0.741	0.163
$0.70\ T$	✗	0.282	**0.920**	**0.769**	**0.140**
$0.80\ T$	✓	0.307	0.876	0.683	0.179

Table 3: Ablation studies. We evaluate the effects of token replacement threshold τ and patch perturbation.

4.4 Applications

As illustrated in Fig. 5, our *Personalize Anything* naturally extends to diverse real-world applications, including subject-driven image generation with layout guidance, inpainting and outpainting. Visualization results in Fig. 2 demonstrate capabilities of our framework on layout-guided personalization, and precise editing with mask conditions, all without architectural modification or fine-tuning.

5 Conclusion

This paper demonstrates that simple token replacement in diffusion transformers (DiTs) enables high-fidelity subject reconstruction by utilizing their position-disentangled representations. The decoupling of semantic features and position allows for the substitution of purely semantic tokens, avoiding positional interference. Building on this, we propose *Personalize Anything*, a training-free image personalization framework that achieves high-fidelity identity preservation without per-subject optimization or large-scale training. Furthermore, our method supports a range of tasks, including layout guidance, multi-subject personalization, and mask-based editing. Leveraging the geometric programming capability of DiTs, our approach enables controllable synthesis and extends naturally to video and 3D generation, advancing scalable customization in generative AI.

Acknowledgments

This work is supported by National Natural Science Foundation of China (62132001), and the Fundamental Research Funds for the Central Universities. And we would like to express our gratitude to our collaborators for their efforts.

References

Cai, S.; Chan, E.; Zhang, Y.; Guibas, L.; Wu, J.; and Wetzstein, G. 2024. Diffusion Self-Distillation for Zero-Shot Customized Image Generation. In *arXiv*.

Chen, X.; Huang, L.; Liu, Y.; Shen, Y.; Zhao, D.; and Zhao, H. 2024a. Anydoor: Zero-shot object-level image customization. In *CVPR*.

Chen, X.; Zhang, Z.; Zhang, H.; Zhou, Y.; Kim, S. Y.; Liu, Q.; Li, Y.; Zhang, J.; Zhao, N.; Wang, Y.; et al. 2024b. UniReal: Universal Image Generation and Editing via Learning Real-world Dynamics. In *arXiv*.

Ding, G.; Zhao, C.; Wang, W.; Yang, Z.; Liu, Z.; Chen, H.; and Shen, C. 2024. FreeCustom: Tuning-Free Customized Image Generation for Multi-Concept Composition. In *CVPR*.

Dosovitskiy, A. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *arXiv*.

Duan, Z.; Ding, Y.; Gou, C.; Zhou, Z.; Smith, E.; and Liu, L. 2024. EZIGen: Enhancing zero-shot subject-driven image generation with precise subject encoding and decoupled guidance. In *arXiv*.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; Podell, D.; Dockhorn, T.; English, Z.; Lacey, K.; Goodwin, A.; Marek, Y.; and Rombach, R. 2024. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In *arXiv*.

Fu, S.; Tamir, N.; Sundaram, S.; Chai, L.; Zhang, R.; Dekel, T.; and Isola, P. 2023. DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. In *arXiv*.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *arXiv*.

Gu, Y.; Wang, X.; Wu, J. Z.; Shi, Y.; Chen, Y.; Fan, Z.; Xiao, W.; Zhao, R.; Chang, S.; Wu, W.; et al. 2024. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In *Advances in Neural Information Processing Systems*.

Han, L.; Li, Y.; Zhang, H.; Milanfar, P.; Metaxas, D.; and Yang, F. 2023. SVDiff: Compact Parameter Space for Diffusion Fine-Tuning. In *arXiv*.

He, J.; Tuo, Y.; Chen, B.; Zhong, C.; Geng, Y.; and Bo, L. 2025. AnyStory: Towards Unified Single and Multiple Subject Personalization in Text-to-Image Generation. In *arXiv*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *arXiv*.

Huang, Q.; Fu, S.; Liu, J.; Jiang, H.; Yu, Y.; and Song, J. 2024. Resolving multi-condition confusion for finetuning-free personalized image generation. In *arXiv*.

Hyung, J.; Shin, J.; and Choo, J. 2024. Magicapture: High-resolution multi-concept portrait customization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2445–2453.

Jiang, J.; Zhang, Y.; Feng, K.; Wu, X.; and Zuo, W. 2024. MC^2: Multi-concept Guidance for Customized Multi-concept Generation. In *arXiv*.

Kong, Z.; Zhang, Y.; Yang, T.; Wang, T.; Zhang, K.; Wu, B.; Chen, G.; Liu, W.; and Luo, W. 2024. Omg: Occlusion-friendly personalized multi-concept generation in diffusion models. In *ECCV*.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *CVPR*.

Kwon, G.; Jenni, S.; Li, D.; Lee, J.-Y.; Ye, J. C.; and Heilbron, F. C. 2024. Concept Weaver: Enabling Multi-Concept Fusion in Text-to-Image Models. In *CVPR*.

Labs, B. F. 2024. FLUX. [Online]. https://github.com/black-forest-labs/flux.

Le, D. H.; Pham, T.; Lee, S.; Clark, C.; Kembhavi, A.; Mandt, S.; Krishna, R.; and Lu, J. 2024. One Diffusion to Generate Them All. In *arXiv*.

Li, D.; Li, J.; and Hoi, S. 2024. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *Advances in Neural Information Processing Systems*.

Li, Z.; Zhang, J.; Lin, Q.; Xiong, J.; Long, Y.; Deng, X.; Zhang, Y.; Liu, X.; Huang, M.; Xiao, Z.; Chen, D.; He, J.; Li, J.; Li, W.; Zhang, C.; Quan, R.; Lu, J.; Huang, J.; Yuan, X.; Zheng, X.; Li, Y.; Zhang, J.; Zhang, C.; Chen, M.; Liu, J.; Fang, Z.; Wang, W.; Xue, J.; Tao, Y.; Zhu, J.; Liu, K.; Lin, S.; Sun, Y.; Li, Y.; Wang, D.; Chen, M.; Hu, Z.; Xiao, X.; Chen, Y.; Liu, Y.; Liu, W.; Wang, D.; Yang, Y.; Jiang, J.; and Lu, Q. 2024. Hunyuan-DiT: A Powerful Multi-Resolution Diffusion Transformer with Fine-Grained Chinese Understanding. In *arXiv*.

Liu, M.; She, D.; Pang, J.; Huang, Q.; Ying, J.; He, W.; Hou, Y.; and Fu, S. 2025. TFCustom: Customized Image Generation with Time-Aware Frequency Feature Guidance. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*.

Liu, Z.; Zhang, Y.; Shen, Y.; Zheng, K.; Zhu, K.; Feng, R.; Liu, Y.; Zhao, D.; Zhou, J.; and Cao, Y. 2023. Cones 2: Customizable image synthesis with multiple subjects. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*.

Ma, J.; Liang, J.; Chen, C.; and Lu, H. 2024. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*.

Ni, Y.; Wen, S.; Koniusz, P.; and Cherian, A. 2025. Noise Consistency Regularization for Improved Subject-Driven Image Synthesis. In *arXiv*.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y.; Li, S.-W.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2024. DINOv2: Learning Robust Visual Features without Supervision. In *arXiv*.

Patel, M.; Jung, S.; Baral, C.; and Yang, Y. 2024. λ-ECLIPSE: Multi-Concept Personalized Text-to-Image Diffusion Models by Leveraging CLIP Latent Space. In *arXiv*.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *arXiv*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *arXiv*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. In *arXiv*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *arXiv*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *arXiv*.

Rout, L.; Chen, Y.; Ruiz, N.; Caramanis, C.; Shakkottai, S.; and Chu, W.-S. 2024. Semantic image inversion and editing using rectified stochastic differential equations. In *arXiv*.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in neural information processing systems*.

Shah, V.; Ruiz, N.; Cole, F.; Lu, E.; Lazebnik, S.; Li, Y.; and Jampani, V. 2024. Ziplora: Any subject in any style by effectively merging loras. In *ECCV*.

Shin, C.; Choi, J.; Kim, H.; and Yoon, S. 2024. Large-Scale Text-to-Image Model with Inpainting is a Zero-Shot Subject-Driven Image Generator. In *arXiv*.

Song, J.; Meng, C.; and Ermon, S. 2022. Denoising Diffusion Implicit Models. In *arXiv*.

Song, K.; Zhu, Y.; Liu, B.; Yan, Q.; Elgammal, A.; and Yang, X. 2024. Moma: Multimodal llm adapter for fast personalized image generation. In *ECCV*.

Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; and Liu, Y. 2024. Roformer: Enhanced transformer with rotary position embedding. In *Neurocomputing*.

Tan, Z.; Liu, S.; Yang, X.; Xue, Q.; and Wang, X. 2024. Ominicontrol: Minimal and universal control for diffusion transformer. In *arXiv*.

Tao, J.; Zhang, Y.; Wang, Q.; Cheng, Y.; Wang, H.; Bai, X.; Zhou, Z.; Li, R.; Wang, L.; Wang, C.; Lin, Q.; and Lu, Q. 2025. InstantCharacter: Personalize Any Characters with a Scalable Diffusion Transformer Framework. In *arXiv*.

Tewel, Y.; Gal, R.; Chechik, G.; and Atzmon, Y. 2024a. Key-Locked Rank One Editing for Text-to-Image Personalization. In *arXiv*.

Tewel, Y.; Kaduri, O.; Gal, R.; Kasten, Y.; Wolf, L.; Chechik, G.; and Atzmon, Y. 2024b. Training-free consistent text-to-image generation. In *ACM Transactions on Graphics (TOG)*.

Voynov, A.; Chu, Q.; Cohen-Or, D.; and Aberman, K. 2023. P+: Extended Textual Conditioning in Text-to-Image Generation. In *arXiv*.

Wang, J.; Li, X.; Zhang, J.; Xu, Q.; Zhou, Q.; Yu, Q.; Sheng, L.; and Xu, D. 2023. Diffusion model is secretly a training-free open vocabulary semantic segmenter. In *arXiv*.

Wang, X.; Fu, S.; Huang, Q.; He, W.; and Jiang, H. 2025. MS-Diffusion: Multi-subject Zero-shot Image Personalization with Layout Guidance. In *arXiv*.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. In *arXiv*.

Zhang, Y.; Song, Y.; Liu, J.; Wang, R.; Yu, J.; Tang, H.; Li, H.; Tang, X.; Hu, Y.; Pan, H.; et al. 2024. Ssr-encoder: Encoding selective subject representation for subject-driven generation. In *CVPR*.

Zhou, Y.; Zhou, D.; Cheng, M.-M.; Feng, J.; and Hou, Q. 2024. StoryDiffusion: Consistent Self-Attention for Long-Range Image and Video Generation. In *arXiv*.

Zhu, C.; Li, K.; Ma, Y.; He, C.; and Xiu, L. 2024. MultiBooth: Towards Generating All Your Concepts in an Image from Text. In *arXiv*.

Zong, Z.; Jiang, D.; Ma, B.; Song, G.; Shao, H.; Shen, D.; Liu, Y.; and Li, H. 2024. Easyref: Omni-generalized group image reference for diffusion models via multimodal llm. In *arXiv*.