



PDF Download
3757377.3763820.pdf
02 March 2026
Total Citations: 1
Total Downloads: 732

Latest updates: <https://dl.acm.org/doi/10.1145/3757377.3763820>

RESEARCH-ARTICLE

In-Context Brush: Zero-shot Customized Subject Insertion with Context-Aware Latent Space Manipulation

YU XU, Institute of Computing Technology Chinese Academy of Sciences, Beijing, Beijing, China

FAN TANG, Institute of Computing Technology Chinese Academy of Sciences, Beijing, Beijing, China

YOU WU, Institute of Computing Technology Chinese Academy of Sciences, Beijing, Beijing, China

LIN GAO, Institute of Computing Technology Chinese Academy of Sciences, Beijing, Beijing, China

OLIVER DEUSSEN, University of Konstanz, Konstanz, Baden-Wurttemberg, Germany

HONGBIN YAN, University of Chinese Academy of Sciences, Beijing, China

[View all](#)

Open Access Support provided by:

Institute of Computing Technology Chinese Academy of Sciences

University of Konstanz

University of Chinese Academy of Sciences

National Cheng Kung University

Published: 14 December 2025

[Citation in BibTeX format](#)

SA Conference Papers '25: SIGGRAPH
Asia 2025 Conference Papers
December 15 - 18, 2025
Hong Kong, Hong Kong

Conference Sponsors:
SIGGRAPH

In-Context Brush: Zero-shot Customized Subject Insertion with Context-Aware Latent Space Manipulation

YU XU, Institute of Computing Technology, Chinese Academy of Sciences, China
 FAN TANG, Institute of Computing Technology, Chinese Academy of Sciences, China
 YOU WU, Institute of Computing Technology, Chinese Academy of Sciences, China
 LIN GAO, Institute of Computing Technology, Chinese Academy of Sciences, China
 OLIVER DEUSSEN, University of Konstanz, Germany
 HONGBIN YAN, University of Chinese Academy of Sciences, China
 JINTAO LI, Institute of Computing Technology, Chinese Academy of Sciences, China
 JUAN CAO*, Institute of Computing Technology, Chinese Academy of Sciences, China
 TONG-YEE LEE, National Cheng-Kung University, Taiwan



Fig. 1. Our method achieves identity-preserving subject insertion in the novel scene harmoniously, simultaneously enabling diverse text-driven control.

*Corresponding author: Juan Cao.

Authors' Contact Information: Yu Xu, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, xuyu21b@ict.ac.cn; Fan Tang, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, f.tan.108@gmail.com; You Wu, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, wuyou22s@ict.ac.cn; Lin Gao, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, gaolin@ict.ac.cn; Oliver Deussen, University of Konstanz, Konstanz, Germany, oliver.deussen@uni-konstanz.de; Hongbin Yan, University of Chinese Academy of Sciences, Beijing, China, yanhongbin24@mails.uicas.ac.cn; Jintao Li, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, jtli@ict.ac.cn; Juan Cao, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, caojuan@ict.ac.cn; Tong-Yee Lee, National Cheng-Kung University, Tainan, Taiwan, tonylee@mail.ncku.edu.tw.



This work is licensed under a Creative Commons Attribution 4.0 International License. SA Conference Papers '25, Hong Kong, Hong Kong © 2025 Copyright held by the owner/author(s). ACM ISBN 979-8-4007-2137-3/25/12 https://doi.org/10.1145/3757377.3763820

Recent advances in diffusion models have enhanced multimodal-guided visual generation, enabling customized subject insertion that seamlessly “brushes” user-specified objects into a given image guided by textual prompts. However, existing methods often struggle to insert customized subjects with high fidelity and align results with the user’s intent through textual prompts. In this work, we propose **In-Context Brush**, a zero-shot framework for customized subject insertion by reformulating the task within the paradigm of in-context learning. Without loss of generality, we formulate the object image and the textual prompts as cross-modal demonstrations, and the target image with the masked region as the query. The goal is to inpaint the target image with the subject aligning textual prompts without model tuning. Building upon a pretrained MMDiT-based inpainting network, we perform test-time enhancement via dual-level latent space manipulation: intra-head *latent feature shifting* within each attention head that dynamically shifts attention outputs to reflect the desired subject semantics and inter-head *attention reweighting* across different heads that amplifies prompt controllability through differential attention prioritization. Extensive experiments and applications demonstrate that our approach achieves superior identity

preservation, text alignment, and image quality compared to existing state-of-the-art methods, without requiring dedicated training or additional data collection. Project page: <https://yuci-gpt.github.io/In-Context-Brush/>.

CCS Concepts: • **Computing methodologies** → **Image manipulation**; *Image processing*.

Additional Key Words and Phrases: customized generation, subject insertion, diffusion models

ACM Reference Format:

Yu Xu, Fan Tang, You Wu, Lin Gao, Oliver Deussen, Hongbin Yan, Jintao Li, Juan Cao, and Tong-Yee Lee. 2025. In-Context Brush: Zero-shot Customized Subject Insertion with Context-Aware Latent Space Manipulation. In *SIG-GRAPH Asia 2025 Conference Papers (SA Conference Papers '25)*, December 15–18, 2025, Hong Kong, Hong Kong. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3757377.3763820>

1 Introduction

Image customization [Gal et al. 2023; Ruiz et al. 2023; Xu et al. 2025], where users aim to render specific subjects into new contexts, has received increasing attention with the advancement of text-to-image diffusion models [Esser et al. 2024; Peebles and Xie 2023; Podell et al. 2024; Rombach et al. 2022]. Beyond synthesizing new scenes from scratch, a more practical and challenging task is to insert a customized subject into a specific region of existing images. This task requires maintaining high semantic fidelity to the customized subject, ensuring contextual harmony with the background, and enabling flexible contextual adaptation (e.g., varying pose, attributes, interactions) with textual prompts provided by users.

Initial attempts [Chen et al. 2024a,b; Song et al. 2023, 2024; Yang et al. 2023] for customized subject insertion typically replace text prompts with subject embeddings, allowing visual specification of the subject but inherently limiting the generation under textual guidance. Later efforts [Choi et al. 2023; Gu et al. 2024a,b; Li et al. 2023] adopt a more straightforward way of learning subjects by fine-tuning the model, and then inserting them into target scenes via additional editing modules. However, such a workflow suffers from subject overfitting and reduced editing controllability. Recent approaches [Li et al. 2024; Wang et al. 2024] share the core objective using techniques, such as inversion and blending, to learn and insert subjects in a training-free manner. However, the low-dimensional latent representations derived from inversion processes inherently restrict textual control precision. Achieving customized subject insertion that harmoniously integrates the subject with visual context (target images) while maintaining identity consistency and adhering to textual context (prompts) with a training-free framework remains challenging to be explored.

Large-scale pre-trained models [Achiam et al. 2023; Touvron et al. 2023] demonstrate remarkable capabilities for context understanding and give rise to in-context learning (ICL) [Brown et al. 2020; Dong et al. 2022; Min et al. 2022], a powerful paradigm that transfers knowledge and facilitates predictions by leveraging input-output pairs, termed as *demonstrations (demos)*, in a zero-shot manner. Similarly, Diffusion Transformers (DiTs) [blackforestlabs.ai 2024; Chen et al. 2024c; Esser et al. 2024; Peebles and Xie 2023] present a promising avenue to incorporate ICL to enable controllable text-to-image generation by utilizing text-image pairs as *demos* and

vision/language conditions as *queries*, generating images that incorporate information from demos while following the specified conditions [Zeng et al. 2024].

However, current ICL-based image generation methods [Najdenkoska et al. 2024; Wang et al. 2023b,a] primarily focus on shallow task adaptation of image-text correspondences in demonstration pairs (e.g., pixel-to-caption matching) while failing to disentangle and transfer abstract subject semantics (e.g., cross-demo categorical invariants or relational patterns). Furthermore, these task-specific conditioning mechanisms conflate subject identity with environmental context, thereby constraining zero-shot generalization to novel subject-scene combinations. As a result, directly leveraging existing ICL frameworks for customized image editing remains a significant challenge.

In this paper, we dive into the ICL framework to enable zero-shot subject insertion. Subject images and textual prompts serve as *demos*, while target images act as *queries* for conducting regional insertion. Following the ICL paradigm, where demos and queries are concatenated as input, we also concatenate prompt tokens and subject image tokens with target image tokens in DiTs to construct an ICL-based inpainting framework. With this framework, we formulate fine-grained subject-level transfer as shifting hidden states in DiTs and propose an intra-head *latent feature shift* injection mechanism to incorporate hidden states of subjects and textual prompts into queries. This enables customized subject-level injection, maintaining consistency between subject and output images while aligning with textual prompts. Additionally, we introduce inter-head attention activation to improve textual control to subjects according to various prompts, and token blending to improve consistency between the inserted subject and the background. Experiments on benchmark datasets show that our method successfully inserts customized subjects into new scenes, enables diverse prompt-driven control, preserves subject fidelity, and achieves coherent visual integration. Our contributions can be summarized as follows.

- We propose **In-Context Brush**, a zero-shot customized subject insertion framework that leverages ICL to transfer subject-level features in large-scale text-to-image diffusion models, and achieves superior identity preservation, prompt alignment, and image quality compared to state-of-the-art methods.
- We reformulate subject insertion under the ICL paradigm as a latent feature shifting problem, and introduce a feature shift injection mechanism to enable accurate and consistent transfer of subject semantics into target scenes.
- We further introduce attention head activation for prompts expressiveness enhancement, and propose a token blending strategy to ensure visual coherence between the inserted subject and the surrounding context.

2 Related Work

2.1 In-context learning for image generation

With the scaling of model and dataset sizes, large language models (LLMs) [Achiam et al. 2023; Devlin 2018; Radford et al. 2019; Touvron et al. 2023] have demonstrated remarkable ICL capabilities [Brown et al. 2020; Wei et al. 2022]. ICL enables models to learn

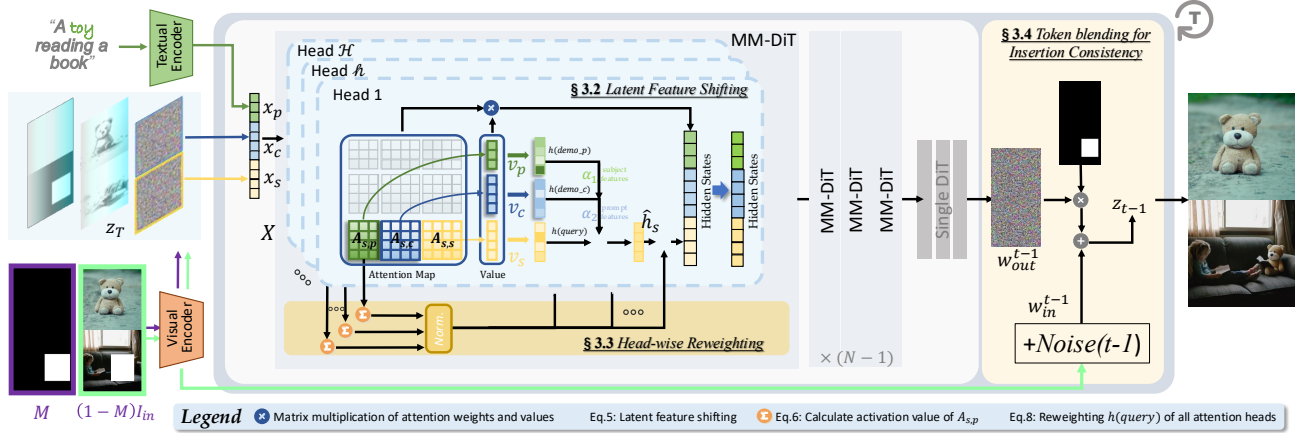


Fig. 2. **Pipeline of our method.** We mainly introduce latent space shifting for subject present in target images in a training-free manner. In the “Latent Feature Shifting” part, features from the reference are shifted to output. We propose attention heads activation for further enhance representation of textual prompts and token blending for consistency injection within the image.

from contextual demonstrations and apply the extracted knowledge to queries. This approach facilitates task execution by conditioning on a combination of demonstrations and query inputs, eliminating the need for parameter optimization. In recent years, the use of ICL has extended beyond natural language processing to encompass image generation. Prompt Diffusion [Wang et al. 2023a] introduces a framework that employs in-context prompts for training across various vision-language tasks, enabling the generation of images from vision-language prompts. Building on this, iPromptDiff [Chen et al. 2023] enhances visual comprehension in visual ICL by decoupling the processing of visual context and image queries while modulating the textual input using integrated context. Furthermore, Context Diffusion [Najdenkoska et al. 2024] separates the encoding of visual context from the query image structure, enabling the model to effectively leverage both visual context and text prompts. However, previous works primarily focus on learning task relationships from demonstrations and transferring them to queries. In contrast, our approach emphasizes learning the semantic feature relationships between subject and target images, enabling subject features insertion into specific regions through a training-free mechanism.

2.2 Customized subject insertion with diffusion models

Previous methods [Chen et al. 2024a,b; Song et al. 2023, 2024; Yang et al. 2023; Yu et al. 2025] typically encode the subject image into embeddings that serve as input conditions to diffusion models. However, text conditions are replaced by image embeddings in the models, making it hard to guide the generation process with prompts. As a result, the output often does not match the user’s intended description, reducing its usefulness. Recent zero-shot approaches [He et al. 2024; Song et al. 2025; Winter et al. 2024b] construct large-scale datasets to train subject insertion models. However, the prompts used in training are typically limited to task-level instructions (e.g., object replacement or removal) or coarse descriptions of the entire scene, which restricts the ability to perform fine-grained control over the inserted subject. Two-stage approaches [Avrahami et al.

2023; Gu et al. 2024a,b; Li et al. 2023; Zhu et al. 2024] first learn subject-specific embeddings through customization techniques [Gal et al. 2023; Ruiz et al. 2023], and then perform insertion into target scenes. While enabling prompt-driven editing, it comes at the cost of subject-specific training, reducing applicability in real-world scenarios. Recently, training-free methods [Li et al. 2024; Lu et al. 2023; Wang et al. 2024; Zhang et al. 2025] have emerged to avoid tuning. These approaches perform inversion of both the subject and scene images into the diffusion latent space, then combine them via a training-free mechanism. However, these methods provide limited controllability through textual prompts due to the lack of explicit alignment between subject semantics and prompt guidance. StyleAligned [Hertz et al. 2024] and ConsiStory [Tewel et al. 2024] explore the feature sharing between reference and target images for stylization and consistency generation tasks, While StyleAligned focuses on stylistic control, it lacks structural precision, limiting its use for subject insertion. ConsiStory ensures image consistency but struggles to preserve identity when learning from given images. ObjectDrop [Winter et al. 2024a], OmniGen [Xiao et al. 2025], and UniReal [Chen et al. 2025] also leverage in-context learning for object insertion or customized generation and show promising results following a training-based fashion, whereas our method operates purely at inference time. IC-LoRA [Huang et al. 2024] activates the in-context generation capabilities of DiTs by training task-specific LoRA modules using paired datasets, but its reliance on data collection and retraining limits practicality. In contrast, our method is training-free and uses ICL to transfer subject features across tasks efficiently. A concurrent work, Diptych Prompting [Shin et al. 2024] also leverages attention in a training-free manner. However, it re-weights attention to emphasize reference influence, which may overlook subject relationships and cause identity inconsistency. In contrast, our method integrates visual features and textual guidance via ICL, achieving stronger alignment with prompts while preserving subject identity.

3 Method

Given a subject image $I_c \in \mathbb{R}^{H \times W \times 3}$ containing the subject to be inserted, a target image $I_s \in \mathbb{R}^{H \times W \times 3}$ providing the background context, a textual prompt p describing the desired output subject, and a binary mask $m \in \{0, 1\}^{H \times W}$ specifying the insertion region, we aim to transfer the subject from I_c into the mask region of I_s with guidance from user provided p , and get final output image I_{gen} . To do so, in Sec. 3.1, we formulate customized image insertion with ICL in DiTs. In Sec. 3.2, we introduce our core mechanism, latent feature shifting, which enables subject transfer in latent space. In Sec. 3.3, we present head-wise reweighting to enhance textual control. In Sec. 3.4, we describe token blending, which ensures better visual consistency between the inserted subject and the background.

3.1 Preliminary

We adopt multi-modal diffusion transformers (MM-DiT) [blackforestlabs.ai 2024; Esser et al. 2024] as the backbone of our generation framework. In each sampling step, MM-DiT take a combination of text and image token embeddings as input and progressively denoise a latent representation to synthesize the output image. To integrate customized subject insertion into MM-DiT, we introduce an ICL paradigm to model the subject-level relationships.

In the setting of ICL in LLMs, consider the translation task, given a few demo prompts, the model will infer task rules based on this task-wise contextual information and translate new input queries. In our scenario, we propose a feature-wise ICL paradigm instead, which transfer subject feature from demo to query. Specifically, to utilize ICL, we construct input demonstrations analogous to those in large language models: the prompt p and subject image I_c jointly serve as *demonstration (demo)*, providing contextual information, while the target image I_s is the *query* whose corresponding region will be inserted. Formally, we concatenate I_c and I_s into a single input image $I_{in} = [I_c; I_s]$, and the mask is correspondingly extended as $M = [0; m]$. In this ICL-based configuration, MM-DiT implicitly learns to transfer subject-level features from the demo (p, I_c) into the query image (I_s) by latent space shifting, detailed in Sec. 3.2. To precisely insert the subject into the background image, we additionally apply Grounding DINO [Liu et al. 2025] and Segment Anything Model (SAM) [Kirillov et al. 2023] to remove the original background in I_c , isolating the desired subject clearly. As a result, with a generation model G_θ , the output image $I_{gen} \in \mathbb{R}^{H \times W \times 3}$ can be formally predicted as:

$$\begin{aligned} [I_c; I_{gen}] &= G_\theta(p, I_{in}, M), \\ &= G_\theta(p, [I_c; I_s], [0_{H \times W}; m]). \end{aligned} \quad (1)$$

3.2 Latent feature shifting for subject injection

In this section, we prove that subject-level features can be injected by shifting hidden states within the framework of ICL, effectively leveraging information from multi-modal demos. In Sec. 3.1, p and I_c are concatenated within attention blocks and used to compute the final hidden states through a joint-attention mechanism. Specifically, let $X = \text{Concatenate}([x_p, x_c, x_s])$ represent the input embedding, where x_p, x_c and x_s represent input token embeddings at the same concatenating positions as p, I_c and I_s , respectively. Let W_q, W_k , and W_v be the learnable key, query, and value matrices for computing the

attention features Q, K , and V , the output hidden states of attention blocks can be formulated as:

$$\hat{h} = \text{Attn}(XW_q, XW_k, XW_v) = \text{Concatenate}([h_p, h_c, h_s]), \quad (2)$$

where h_p, h_c, h_s represent the hidden states corresponding to each component in X . We put the detailed derivation in supplementary materials.

Although in attention blocks, the overall feature X is processed in a self-attention manner, there also exist relationships in the form of cross-attention among different pairs of its components. For example, h_s is directly composed of two parts: one part is derived from the self-attention computation of x_s ; the other part is obtained through the interaction with features provided by the textual prompt and the reference subject, i.e., x_p and x_c . We only focus on h_s because the generated result I_{gen} is directly related to it.

This characteristic activates us to leverage contextual information from other features in the latent space from the perspective of in-context learning. When x_p and x_c interact through cross-attention with x_s respectively, they serve as demo providing semantic feature-wise contextual information and generating the attention output $h(\text{demo}_p)$ and $h(\text{demo}_c)$. As for the position for insertion, the self-attention computation of x_s itself yields the original output $h(\text{query})$ without demo. Therefore, we rewrite the formula of h_s in the form of the attention operation:

$$\begin{aligned} h_s &= \text{Softmax} \left(\begin{bmatrix} x_s W_{qk} x_p^\top & x_s W_{qk} x_c^\top & x_s W_{qk} x_s^\top \end{bmatrix} \right) \begin{bmatrix} x_p W_v \\ x_c W_v \\ x_s W_v \end{bmatrix} \\ &= \alpha_p \cdot h(\text{demo}_p) + \alpha_c \cdot h(\text{demo}_c) + \alpha_s \cdot h(\text{query}), \end{aligned} \quad (3)$$

where $W_{qk} = W_q W_k^\top$. We put the detailed derivation in supplementary materials. α_{tag} is the scalar that represents the sum of normalized attention weights between different hidden states:

$$\alpha_{tag} = \frac{\sum \exp(x_s W_{qk} x_{tag}^\top)}{\sum \exp(x_s W_{qk} x_p^\top) + \sum \exp(x_s W_{qk} x_c^\top) + \sum \exp(x_s W_{qk} x_s^\top)}, \quad (4)$$

where $\alpha_p + \alpha_c + \alpha_s = 1$. Therefore, the essence of this subject-level relationship ICL can be regarded as a latent feature shifting on the original attention output $h(\text{query})$ on the direction figured by $h(\text{demo}_p)$ and $h(\text{demo}_c)$. The attention mechanism of DiTs automatically determines the distance of the shift.

Based on our conclusion, we propose a method named “**feature shift injection**”, a straightforward way that manipulates the shift of attention feature outputs directly related to I_{gen} , to enhance the utilization and focus of DiTs on in-context information from input conditions in customized subject insertion. Specifically, we can divide the weight map for each attention head within the attention blocks into multiple patches, as shown in Fig. 2.

For the convenience of representation, we use $A_{i,j}$ to represent attention map in position of patch $x_i W_{qk} x_j^\top$, and define value feature $V = \text{Concat}([v_p, v_c, v_s]) = \text{Concat}([x_p W_v, x_c W_v, x_s W_v])$. The results of the hidden states h_s are determined solely by the bottom three attention maps $A_{s,p}$, $A_{s,c}$, and $A_{s,s}$. According to Eq. 3, they are respectively computed with the corresponding three parts of the value feature V to obtain $h(\text{demo}_p)$, $h(\text{demo}_c)$, and $h(\text{query})$.

To shift the latent features from $h(query)$, we directly amplify the values of scalars α_p and α_c because they are controlling the influence of $h(demo_p)$ and $h(demo_c)$ on the original latent feature $h(query)$ without demos. In fact, this corresponds to adding the weighted results of separately computing attention maps $A_{s,p}$ and $A_{s,c}$ with v_p and v_c onto the output latent states h_s :

$$\hat{h}_s = h_s + \alpha_1 A_{s,p} v_p + \alpha_2 A_{s,c} v_c, \quad (5)$$

where α_1 and α_2 control the strength of shift like Eq. 3. During inference, the products of $A_{s,p} v_p$ and $A_{s,c} v_c$ are integrated into the attention mechanism as *demo* information. The “shifting” refers to this adjustment process of hidden states using injected *demo* information (which include the features of the subject and the textual prompt) at inference time, enabling capture the subject-level relationships from in-context conditions and generate consistent subjects. This process requires no fine-tuning and operates purely through the injection of *demo* information into the attention computation of the pre-trained model.

3.3 Head-wise reweighting for textual control injection

While latent feature shifting mechanism enables subject transfer, effective control with diverse prompts remains challenging due to strong priors encoded in the reference image. In practice, we observe that inserted subjects often retain undesired attributes (e.g., colors, materials) from the subject image, even when the prompt specifies changes. This common limitation stems from the lack of selective control over semantic attention during generation. To address this, we introduce a head-wise reweighting mechanism that improves the alignment between the generated image and the prompt by adaptively adjusting the contribution of different attention heads. This is motivated by recent findings [Gandelsman et al. 2024; Xu et al. 2024] that attention heads in transformers exhibit semantic specialization—different heads respond to different types of features. Our key insight is to leverage the attention activation in the demo of the ICL setup to estimate which attention heads are most activated by the prompt tokens, and then reweight these heads during generation. As shown in Eq. 3, we leverage $h(demo_p)$ to soft activate $h(query)$ across different attention heads. Specifically, for $h(demo_p)$, we measure the attention maps $A_{p,s}$ across all attention heads and assign different weights to queries based on activation values. The activation value of $A_{p,s}$ in attention head ℓ can be formed as:

$$V_{\ell} = \sum_{i,j} \left(A_{p,s}^{(\ell)} \right)_{i,j}, \quad (6)$$

where i and j are indices of $A_{p,s}^{(\ell)}$. Then we normalize all V_{ℓ} across attention heads following:

$$\hat{V}_{\ell} = \frac{V_{\ell} - \min(V)}{\max(V) - \min(V)}, \ell = 1, 2, \dots, \mathcal{H}, \quad (7)$$

where ℓ is the index of attention heads, and \mathcal{H} is the total number of heads, $\min(V) = \min\{V_1, V_2, \dots, V_{\mathcal{H}}\}$, $\max(V) = \max\{V_1, V_2, \dots, V_{\mathcal{H}}\}$. The final output of hidden states on each attention head are:

$$\hat{h}_{\ell}(query) = h_{\ell}(query) \cdot \hat{V}_{\ell}. \quad (8)$$

This encourages the model to emphasize semantic that are relevant to the user prompt while suppressing prompt-irrelevant heads. This

improves semantic controllability and leads to more faithful editing with respect to user intent.

3.4 Token blending for insertion consistency

In this section, we tackle the challenge of ensuring intra-image consistency when inserting customized subjects by refining feature interactions to mitigate distribution shifts. Specifically, as the subject is injected into a new contextual environment, we further analyze the challenge of ensuring intra-image consistency on customized subject injection.

Due to the semantic differences between the inserted subject and the background, the latent feature of the insertion region within mask m in I_s could be incongruous with the background (i.e., $I_s \cdot (1 - m)$). I_s is expected to guide the inserted subject in x_s to be consistent with target regions in distribution. However, affected by x_c and x_p after each sampling step, the semantic distribution of target region in x_s deviates from the original input after each sampling step. In the next step, the distribution-biased x_s will, in turn, provide erroneous guidance for the subject insertion of the target region (i.e., $x_s \cdot m$) due to the interaction of contextual information in DiTs. Multiple sampling steps will gradually amplify this bias, resulting in an inharmonious fusion effect between the inserted subjects and the background in the result (such as irregular edges or differences in tone).

To prevent the deviation caused by inconsistent distribution in multi-step sampling, we propose effective token blending for insertion consistency. Specifically, suppose the output hidden states of denoising step t is w^t , we add noise to $(1 - M) \cdot I_{in}$ to obtain w_{in}^{t-1} after each step t to $t - 1$. Subsequently, we fuse w_{in}^{t-1} with the output w_{out}^{t-1} of the current step according to the mask M :

$$w_{out}^t = w_{in}^{t-1} + M \cdot w_{out}^{t-1}. \quad (9)$$

By this method, we ensure that in each step, the inserted region can be correctly guided by unbiased background semantics in distribution, thus enhancing the consistency between the inserted subjects and the context of the background in I_{gen} .

4 Experiments

In this section, we first introduce the experiment settings, and present qualitative and quantitative results in Sec. 4.2 and Sec. 4.3. We further evaluate two-stage methods that combine subject insertion methods with customization methods or editing methods for a comprehensive comparison. Finally, we conduct an ablation study on hyperparameters and proposed modules. For more details on multiple seeds and time cost, please refer to supplementary materials.

4.1 Experiments settings

Baselines. We compare our method with eight state-of-the-art text and image-guided image generation methods, including training-based methods Break-a-scene [Avrahami et al. 2023], Swap-anything [Gu et al. 2024b], DreamEdit [Li et al. 2023], IC-LoRA [Huang et al. 2024], and training-free methods TF-ICON [Lu et al. 2023], TIGIC [Li

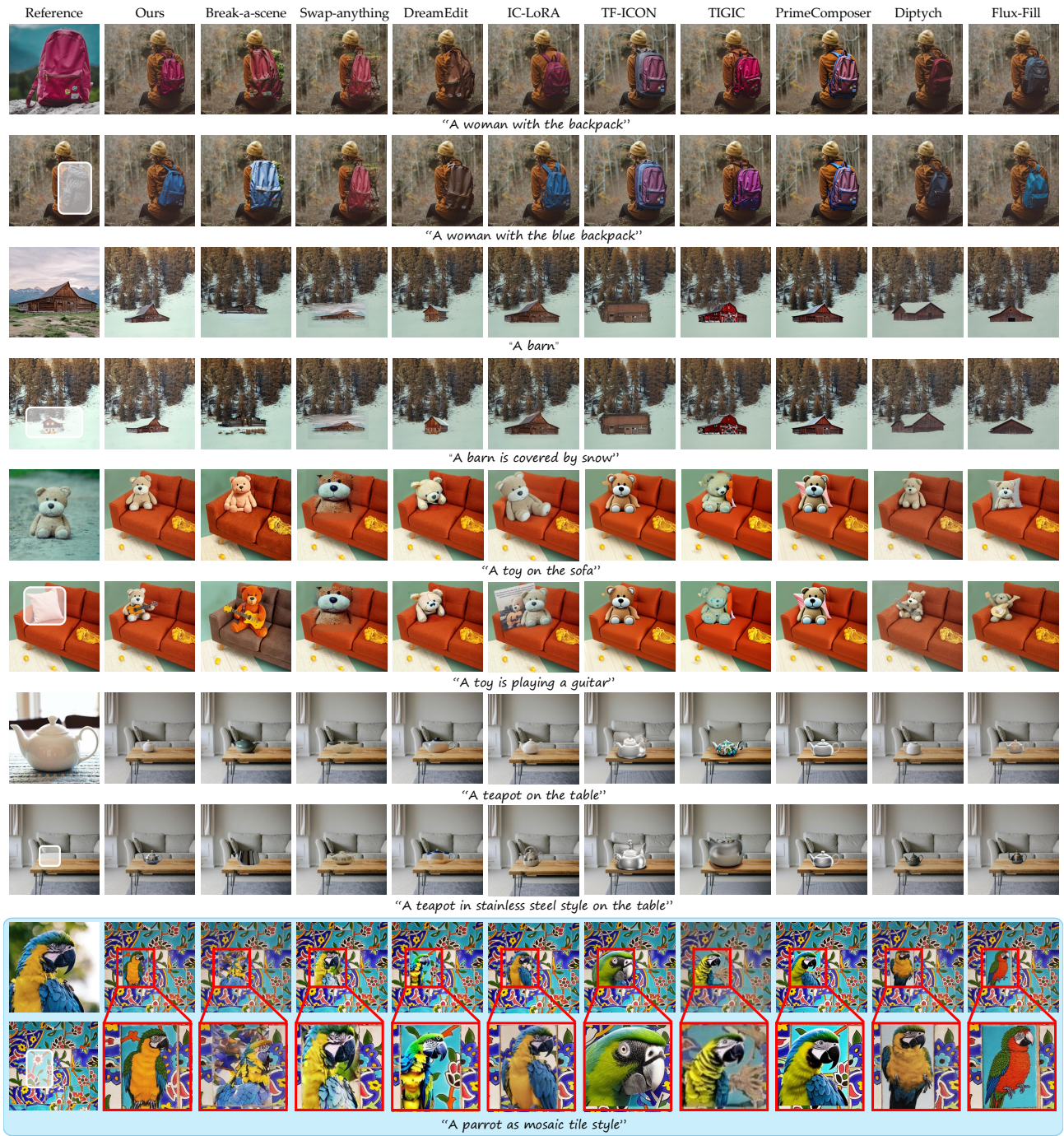


Fig. 3. **Qualitative comparison on subject injection and editing with baseline methods.** Results of our results maintain identity consistency with reference while preserving fine-grained features, and are also aligning with the prompts. Masks are labeled as white boxes on target images.

Table 1. **Comparison of similarity scores between output images and reference images, and between output images and text prompts.** “Injection” evaluates the subject identity alignment between the reference images and the output images. “Editing” evaluates the text alignment between the output images and the corresponding prompts. Our method has the best scores, indicating that our approach successfully edit images guided by text while maintaining consistency with reference images and high image quality.

Methods	DINO(\uparrow)		CLIP-I(\uparrow)		CLIP-T(\uparrow)		FID(\downarrow)
	Injection	Editing	Injection	Editing	Injection	Editing	
Break-a-scene	0.7041 \pm 0.0659	0.7087 \pm 0.0546	0.7128 \pm 0.2076	0.6640 \pm 0.1855	0.2570 \pm 0.0230	0.2820 \pm 0.0659	217.81
Swap-anything	0.7058 \pm 0.0625	0.7035 \pm 0.0556	0.7275 \pm 0.1821	0.7296 \pm 0.1602	0.2493 \pm 0.0181	0.2296 \pm 0.0354	190.69
DreamEdit	0.6521 \pm 0.0625	0.6477 \pm 0.0652	0.7203 \pm 0.1321	0.7169 \pm 0.1377	0.2295 \pm 0.0562	0.2271 \pm 0.0557	176.21
IC-LoRA	0.7005 \pm 0.0631	0.6855 \pm 0.0621	0.6891 \pm 0.1422	0.6511 \pm 0.1325	0.2523 \pm 0.0367	0.2624 \pm 0.0522	149.75
DB+Paint-by-example	0.7037 \pm 0.0601	0.6955 \pm 0.0442	0.7162 \pm 0.0193	0.6841 \pm 0.0167	0.2668 \pm 0.0205	0.2385 \pm 0.0343	172.33
Anydoor+TurboEdit	0.6881 \pm 0.0699	0.6925 \pm 0.0631	0.6823 \pm 0.1401	0.6425 \pm 0.1422	0.2499 \pm 0.0533	0.2499 \pm 0.0565	213.77
MimicBrush+TurboEdit	0.7044 \pm 0.0511	0.6951 \pm 0.0502	0.7107 \pm 0.1337	0.6991 \pm 0.1851	0.2674 \pm 0.0366	0.2551 \pm 0.0621	153.28
TF-ICON	0.7053 \pm 0.0280	0.7061 \pm 0.0225	0.7235 \pm 0.1188	0.7018 \pm 0.1542	0.2334 \pm 0.0270	0.2234 \pm 0.0297	169.04
TIGIC	0.6901 \pm 0.0231	0.6906 \pm 0.0302	0.7145 \pm 0.1744	0.6789 \pm 0.2012	0.2656 \pm 0.0313	0.2272 \pm 0.0631	179.07
PrimeComposer	0.7029 \pm 0.0510	0.6931 \pm 0.0476	0.7124 \pm 0.1128	0.7426 \pm 0.0619	0.2609 \pm 0.0163	0.2300 \pm 0.0532	166.85
Diptych	0.6559 \pm 0.0679	0.6531 \pm 0.0691	0.7225 \pm 0.1287	0.7145 \pm 0.1140	0.2321 \pm 0.0569	0.2287 \pm 0.0403	179.28
Flux-Fill	0.6798 \pm 0.0602	0.6761 \pm 0.0497	0.7668 \pm 0.1274	0.7843 \pm 0.1124	0.2634 \pm 0.0213	0.2667 \pm 0.0288	127.18
Ours w/o head	0.7115 \pm 0.0564	0.6891 \pm 0.0484	0.7882 \pm 0.1201	0.7889 \pm 0.1121	0.2621 \pm 0.186	0.2682 \pm 0.331	123.82
Ours w/o blend	0.7116 \pm 0.0591	0.6902 \pm 0.0519	0.7901 \pm 0.1192	0.7869 \pm 0.1133	0.2633 \pm 0.171	0.2703 \pm 0.294	129.47
Ours	0.7121 \pm 0.0542	0.6945 \pm 0.0563	0.7957 \pm 0.1138	0.7924 \pm 0.1113	0.2685 \pm 0.0173	0.2834 \pm 0.0365	122.61

et al. 2024], PrimeComposer [Wang et al. 2024] and Diptych Prompting [Shin et al. 2024]. We also include two-stage methods, which combine MimicBrush [Chen et al. 2024a] or Anydoor [Chen et al. 2024b] with TurboEdit [Deutch et al. 2024] (first insert subjects to target images then editing images) and combine Dreambooth [Ruiz et al. 2023] with Paint-by-example [Yang et al. 2023] (first learn and edit subjects then inject to target images) for further comparison. IC-LoRA and Diptych Prompting use FLUX.1-dev [blackforestlabs.ai 2024] as the backbone, while all other baselines use Stable Diffusion 2 [Rombach et al. 2022] as the backbone.

Datasets. We collect subject images from Dreambooth datasets [Ruiz et al. 2023], which contains 30 subjects of 15 different classes, and collect 50 diverse scenes as target images from COCO dataset [Lin et al. 2014]. We also collected 50 additional subject images and 80 scene images from the Internet to enable a more diverse and comprehensive evaluation. Thus, the evaluation dataset contains 100 subject images and 130 scene images.

4.2 Qualitative comparisons

We present the visual results compared with customized subject insertion methods in Fig. 3. TF-ICON struggles to maintain semantic information from subject images (in the cases of “backpack”, “barn” and “teapot”) and has difficulty in editing aligning with textual prompts. Break-a-scene has a good ability to follow prompt guidance in most cases. However, it lacks accurate expression of fine-grained features (in the cases of “toy” and “teapot”), and also some obvious artifacts are presented between subject and background (“backpack”, “barn” and “mosaic tile”), leading to overall disharmony. Swap-anything fails to learn the semantic information of the subject, leading to the expression of the subject in the results being close to the copy-move effect and showing limited effects

when editing complex subjects. DreamEdit, TIGIC, and PrimeComposer also fail to accurately edit the subject in accordance with the given prompts, while IC-LoRA fails when handling complex prompts such as “barn”, “toy”, or “mosaic tile”. PrimeComposer also lacks semantic learning ability and generates subjects in copy-move effects (“toy” and “mosaic tile”). Diptych struggles to achieve generate results with high identity alignment with reference subjects (in the case of “backpack”, “barn”, and “teapot”), or editing effects (in the case of “backpack”, “toy”, and “teapot”). Our approach achieves the best identity preservation and prompt-followed editing effects, surpassing the performance of baseline methods.

4.3 Quantitative comparisons

Following previous methods [Gu et al. 2024b; Li et al. 2023; Lu et al. 2023; Song et al. 2024], we evaluate our method in three aspects: subject identity alignment between I_c and I_{gen} , editing alignment between p and I_{gen} , and overall image quality. We use DINO [Liu et al. 2025] and CLIP-I [Ilharco et al. 2021] for subject identity alignment, including subject injection results and subject injection with editing results. We use CLIP-T [Ilharco et al. 2021] to evaluate editing alignment. “Injection” evaluates identity alignment between reference and generated images. “Editing” evaluates prompt alignment. FID [Heusel et al. 2017] is used for evaluating overall image quality. We conduct quantitative comparisons in Tab. 1, displaying the evaluation index results and their standard deviations. As show in Tab. 1, our method outperforms baselines in subject identity alignment, editing alignment, and image quality. Although the editing score (third column) is slightly lower than Break-a-scene and Swap-anything, likely due to their training-based methods that enable more free representation of semantics and are not limited to the reference, our method achieves stronger CLIP-I and CLIP-T scores, indicating higher robustness without requiring additional

training. For fine-grained evaluation of subject fidelity and background preservation, following DreamEdit [Li et al. 2023] and SwapAnything [Gu et al. 2024b], we segment the generated results into subject and target regions. We then compute the similarity between generated and reference images using DINO and CLIP-I features for both regions. These metrics reflect the semantic consistency of the inserted subject and the integrity of the surrounding context. As shown in Tab. 2, our method achieves the highest scores across all four metrics, demonstrating superior subject fidelity preservation and background maintenance.

Table 2. **Quantitative comparison of the subject and background consistency with the subject and target images.** Higher scores in the “Subject” columns indicate better preservation of subject fidelity from the content images, while higher scores in the “Background” columns reflect better preservation of the target image background.

Methods	Subject(↑)		Background(↑)	
	DINO	CLIP-I	DINO	CLIP-I
Break-a-scene	0.8415	0.6315	0.9203	0.7560
Swap-anything	0.8361	0.7329	0.9288	0.7673
DreamEdit	0.8504	0.7429	0.9585	0.7962
IC-LoRA	0.8179	0.7181	0.9333	0.7972
DB+Paint-by-example	0.8224	0.9385	0.9305	0.7851
Anydoor+TurboEdit	0.8502	0.7815	0.9278	0.7698
MimicBrush+TurboEdit	0.8301	0.9297	0.9342	0.7921
TF-ICON	0.8450	0.7632	0.9160	0.7292
TIGIC	0.8483	0.6998	0.9428	0.7954
PrimeComposer	0.8505	0.7725	0.9405	0.7962
Diptych	0.7700	0.6560	0.8734	0.7570
Flux-Fill	0.8405	0.8001	0.9531	0.7853
Ours	0.8523	0.8090	0.9596	0.8100

User study. We conduct a user study to evaluate participant preferences on identity alignment, editing alignment, and overall image quality between our method and the baselines. A total of 65 participants (33 female and 32 male, aged 14 to 55 years) participated in the survey, including 25 researchers specializing in computer graphics or computer vision. Each participant was asked to evaluate 35 cases, resulting in 6,825 votes. We present results in Fig. 4, and from these, we can see that our method achieves the best identity alignment preference. This indicates that the subject identity of the original image is most effectively preserved in the generated image, avoiding distortion. Furthermore, our method receives the highest editing alignment preference, indicating better prompt-driven customization than other approaches, achieving the customized effects desired by users. Overall, users also favor our results for their higher image quality and visual coherence.

4.4 Comparison with two-stage methods

Customized subject insertion can also be achieved through two-stage approaches: by utilizing advanced subject customization techniques [Gal et al. 2023; Ruiz et al. 2023] in the first stage for training custom subject representations and leveraging image-guided editing methods [Yang et al. 2023] in the second stage to inject subjects

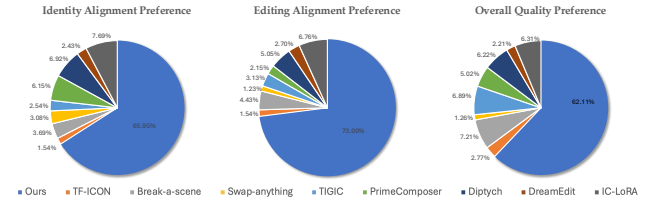


Fig. 4. User study results.

into target images. Also, users can leverage subject injection methods [Chen et al. 2024a,b] in the first stage and utilize text-guided editing methods [Deutch et al. 2024; Hertz et al. 2023] for further subjects editing. We compare both two-stage approaches with our method and present results in Fig. 5. As shown in the figure, Dreambooth with Paint-by-example is challenging to capture fine-grained features, leading to feature and identity inconsistency. Anydoor with TurboEdit loses the ability to preserve subject image identity, while MimicBrush with TurboEdit struggles to follow editing prompts, and the interaction with the background is not harmonious. Our approach achieves the best feature and identity preservation and adapts the generated results to the new scenario with prompt editing, surpassing the performance of two-stage methods.



Fig. 5. Comparisons with two-stage methods.

4.5 Ablation study

Shift strength of α . In Eq. 5, α_1 and α_2 are shift strength parameters for controlling guidance strength from textual prompts and subject images. We further examine the impact of different values for α_1 and α_2 on the generation results. When testing α_1 , α_2 is set to 0.5, and vice versa. The results, as shown in Fig. 6, indicate that as α_1 increases, the expressiveness of the textual prompt in the generated result gradually strengthens. Similarly, increasing α_2 enhances the image expression, causing the identity of the generated subject to align more closely with the reference. However, excessively large values of α_1 and α_2 (e.g., 0.5) lead to a decrease in image quality. Therefore, users can adjust these parameters based on the specific image to balance the control of the image and generation quality.

Customized subject insertion via basic Flux-Fill model. We build an inpainting pipeline with I_c and I_s concatenated as input to evaluate the basic customized subject insertion ability of Flux-Fill. Results in

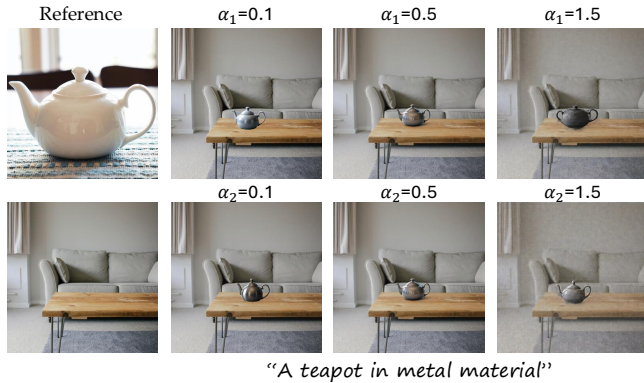


Fig. 6. Ablation study on shift strength.



Fig. 7. Ablation study on token blending.

Fig. 3 show that, to some extent, the basic pipeline generates subjects similar to I_c (although the details are not aligned enough with the reference image); however, when editing the subjects, identity consistency is significantly reduced, indicating that without our methods, the basic model does not learn semantic level information, resulting in limited editing ability.

Without token blending. We ablate the token blending module and present results in Fig. 7. We can see that with the introduction of latents from the target and fusion with latents from subjects along the denoising step, the presentation of subjects in the target images has better interaction with the background, achieving the overall consistency of the image (e.g., the subject and background hue in the image are consistent in the first case and the toy has better interaction with broom when editing it as sitting on the broom in the second case). We also present quantitative results in Tab. 1, and results show that without token blending, the FID score increases, indicating the overall image quality decreases. The CLIP-T score decreases, indicating that the results have a lower alignment with the prompts. Without token blending, subjects have less interaction with background, leading to less following the prompts.

Without attention heads reweighting. We evaluate prompt representation by ablating attention heads reweighting method, reweighting all heads equally, and present results in Fig. 8. The findings

reveal that, in the first case, due to the prior influence of images (white ceramic material), it is difficult to effectively edit the teapot (metal materials) without reweighting key heads. In the second case, although the text expresses “white clothes”, the outputs are still affected by the reference and generate pink clothes, which fails to reflect the intended prompt. Quantitative results in Tab. 1 show that without reweighting, the FID score increases, indicating lower image quality. Furthermore, the decrease in CLIP-T score also shows reduced alignment with the editing prompt. Overall, without head-wise enhancement for prompt representation, the generated results are greatly influenced by references and difficult to control by the prompt, leading to sub-optimal generation effects.



Fig. 8. Ablation study on attention heads activation.

5 Applications

Virtual try on. A key application of our method is Virtual Try-On (VTON), which involves digitally dressing a target person with specified clothing. This is widely used in fashion retail to help users visualize outfits before purchase. As shown in Fig. 9, our method accurately transfers garments to the target subject while preserving identity and achieving strong alignment with user prompts. It handles various clothing types and styles, demonstrating versatility for real-world fashion scenarios.

Compositional generation. Another application of our method is compositional generation, where users iteratively insert multiple elements into a scene with layout control via masks. This is especially useful in tasks like interior design, enabling users to explore combinations of furniture by adjusting placements and styles. As shown in Fig. 10, our framework supports flexible and coherent scene construction, allowing users to refine designs interactively and visualize personalized arrangements with ease.

Partly insertion. Our method supports selective part-based insertion, enabling users to transfer specific regions from reference into corresponding locations of the generated output. This allows fine-grained control while maintaining spatial and semantic alignment. As shown in Fig. 11, we insert the wheels of a reference car and the legs of chairs into target scenes. In both cases, the inserted parts preserve high fidelity and blend naturally with the background, demonstrating the method’s effectiveness for precise partial edits in applications such as product variation and scene refinement.

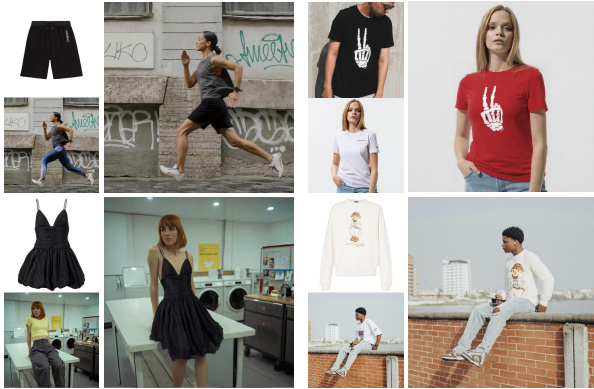


Fig. 9. Application of virtual try-on.



Fig. 10. Application of compositional generation.

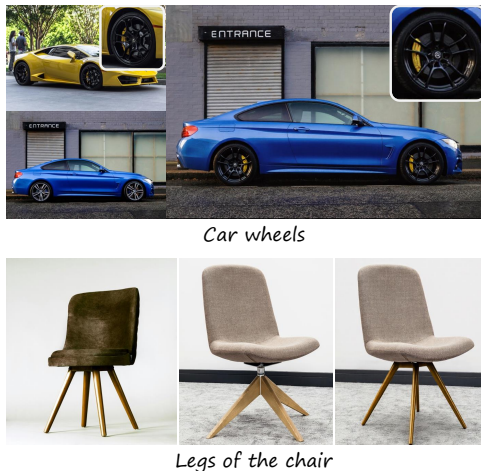


Fig. 11. Application of partly insertion.

6 Limitations and badcase

When target images contain subjects similar to those in the subject images, the generated results may exhibit features resembling the target image. For example, as shown in Fig. 12, certain patterns on the windows of the generated vehicle are similar to the corresponding positions in the target image’s background. This issue likely arises because similar contextual features are referenced during the

Fig. 12. **Badcase.** In some cases, interaction among similar contextual features in attention calculation may cause same features in appearance of results as the concepts from background.

self-attention calculation. To address this in future work, we can consider introducing constraints on the attention mechanism.

7 Conclusions

In this work, we leverage ICL to activate the context-consistent generation capability of large-scale pre-trained text-to-image models, enabling customized subject insertion. By reformulating ICL as latent space shifting, we achieve zero-shot insertion of specific subjects into novel images. Additionally, we employ head-wise reweighting and token blending to enhance the insertion consistency of text attribute expression. Extensive quantitative and qualitative experiments and user study demonstrate the superiority of our approach over existing state-of-the-art methods.

Acknowledgments

We thank the reviewers for their insightful comments and suggestions, as well as Yuxin Zhang for the discussion. This work was partly supported by Beijing Science and Technology Plan Project under No. Z231100005923033, National Science and Technology Council, Taiwan under Grant 114-2221-E-006-114-MY3, and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy–EXC 2117–422037984.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. 2023. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*. 1–12.
- blackforestlabs.ai. 2024. FLUX, offering state-of-the-art performance image generation. <https://blackforestlabs.ai/>. Accessed: 2024-10-07.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- Junsong Chen, Jincheng YU, Chongjian GE, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024c. PixArt- α : Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=eAKmQP3m1>
- Tianqi Chen, Yongfei Liu, Zhendong Wang, Jianbo Yuan, Quanzeng You, Hongxia Yang, and Mingyuan Zhou. 2023. Improving in-context learning in diffusion models with visual context-modulated prompts. *arXiv preprint arXiv:2312.01408* (2023).
- Xi Chen, Yutong Feng, Mengting Chen, Yiyang Wang, Shilong Zhang, Yu Liu, Yujun Shen, and Hengshuang Zhao. 2024a. Zero-shot image editing with reference imitation. *Advances in Neural Information Processing Systems* 37 (2024), 84010–84032.
- Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. 2024b. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6593–6602.

- Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. 2025. Unreal: Universal image generation and editing via learning real-world dynamics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 12501–12511.
- Jooyoung Choi, Yunjey Choi, Yunji Kim, Junho Kim, and Sungroh Yoon. 2023. Customedit: Text-guided image editing with customized diffusion models. *arXiv preprint arXiv:2305.15779* (2023).
- Gilad Deutch, Rinon Gal, Daniel Garibi, Or Patashnik, and Daniel Cohen-Or. 2024. Turboedit: Text-based image editing using few-step diffusion models. In *SIGGRAPH Asia 2024 Conference Papers*. 1–12.
- Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234* (2022).
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*.
- Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. 2023. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *The Eleventh International Conference on Learning Representations*.
- Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. 2024. Interpreting CLIP’s Image Representation via Text-Based Decomposition. In *The Twelfth International Conference on Learning Representations*.
- Jing Gu, Yilin Wang, Nanxuan Zhao, Tsu-Jui Fu, Wei Xiong, Qing Liu, Zhifei Zhang, He Zhang, Jianming Zhang, HyunJoon Jung, et al. 2024a. Photoswap: Personalized subject swapping in images. *Advances in Neural Information Processing Systems* 36 (2024).
- Jing Gu, Nanxuan Zhao, Wei Xiong, Qing Liu, Zhifei Zhang, He Zhang, Jianming Zhang, HyunJoon Jung, Yilin Wang, and Xin Eric Wang. 2024b. SwapAnything: Enabling Arbitrary Object Swapping in Personalized Image Editing. In *European Conference on Computer Vision*. 402–418.
- Runze He, Kai Ma, Linjiang Huang, Shaofei Huang, Jialin Gao, Xiaoming Wei, Jiao Dai, Jizhong Han, and Si Liu. 2024. Freeedit: Mask-free reference-based image editing with multi-modal instruction. *arXiv preprint arXiv:2409.18071* (2024).
- Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. 2023. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *The Eleventh International Conference on Learning Representations*.
- Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. 2024. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4775–4785.
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).
- Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. 2024. In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775* (2024).
- Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. *OpenCLIP*. doi:10.5281/zenodo.5143773
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4026.
- Pengzhi Li, Qiang Nie, Ying Chen, Xi Jiang, Kai Wu, Yuhuan Lin, Yong Liu, Jinlong Peng, Chengjie Wang, and Feng Zheng. 2024. Tuning-free image customization with image and text guidance. In *European Conference on Computer Vision*. Springer, 233–250.
- Tianle Li, Max Ku, Cong Wei, and Wenhu Chen. 2023. DreamEdit: Subject-driven Image Editing. *Transactions on Machine Learning Research* (2023).
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Springer, 740–755.
- Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. 2025. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*. Springer, 38–55.
- Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. 2023. Tf-icon: Diffusion-based training-free cross-domain image composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2294–2305.
- Sewon Min, Xixi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?. In *EMNLP*.
- Ivona Najdenkoska, Animesh Sinha, Abhimanyu Dubey, Dhruv Mahajan, Vignesh Ramanathan, and Filip Radenovic. 2024. Context diffusion: In-context aware image generation. In *European Conference on Computer Vision*. Springer, 375–391.
- William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4195–4205.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on Learning Representations*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. (2019).
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
- Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. 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*. 22500–22510.
- Chaehun Shin, Jooyoung Choi, Heeseung Kim, and Sungroh Yoon. 2024. Large-Scale Text-to-Image Model with Inpainting is a Zero-Shot Subject-Driven Image Generator. *arXiv preprint arXiv:2411.15466* (2024).
- Wensong Song, Hong Jiang, Zongxing Yang, Ruijie Quan, and Yi Yang. 2025. Insert Anything: Image Insertion via In-Context Editing in DiT. *arXiv preprint arXiv:2504.15009* (2025).
- Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. 2023. Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18310–18319.
- Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, He Zhang, Wei Xiong, and Daniel Aliaga. 2024. Imprint: Generative object compositing by learning identity-preserving representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8048–8058.
- Yoad Twel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. 2024. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)* 43, 4 (2024), 1–18.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. 2023b. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6830–6839.
- Yibin Wang, Weizhong Zhang, Jianwei Zheng, and Cheng Jin. 2024. Primecomposer: Faster progressively combined diffusion for image composition with attention steering. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 10824–10832.
- Zhendong Wang, Yifan Jiang, Yadong Lu, Pengcheng He, Weizhu Chen, Zhangyang Wang, Mingyuan Zhou, et al. 2023a. In-context learning unlocked for diffusion models. *Advances in Neural Information Processing Systems* 36 (2023), 8542–8562.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research* (2022).
- Daniel Winter, Matan Cohen, Shlomi Fruchter, Yael Pritch, Alex Rav-Acha, and Yedid Hoshen. 2024a. Objectdrop: Bootstrapping counterfactuals for photorealistic object removal and insertion. In *European Conference on Computer Vision*. Springer, 112–129.
- Daniel Winter, Asaf Shul, Matan Cohen, Dana Berman, Yael Pritch, Alex Rav-Acha, and Yedid Hoshen. 2024b. ObjectMate: A Recurrence Prior for Object Insertion and Subject-Driven Generation. *arXiv preprint arXiv:2412.08645* (2024).
- Shitao Xiao, Yuezhe Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. 2025. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 13294–13304.
- Yu Xu, Fan Tang, Juan Cao, Yuxin Zhang, Oliver Deussen, Weiming Dong, Jintao Li, and Tong-Yee Lee. 2025. B4M: Breaking Low-Rank Adapter for Making Content-Style Customization. *ACM Transactions on Graphics* 44, 2 (2025), 1–17.
- Yu Xu, Fan Tang, Juan Cao, Yuxin Zhang, Xiaoyu Kong, Jintao Li, Oliver Deussen, and Tong-Yee Lee. 2024. HeadRouter: A Training-free Image Editing Framework for MM-DITs by Adaptively Routing Attention Heads. *arXiv preprint arXiv:2411.15034* (2024).
- Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. 2023. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18381–18391.

- Yongsheng Yu, Ziyun Zeng, Haitian Zheng, and Jiebo Luo. 2025. OmniPaint: Mastering Object-Oriented Editing via Disentangled Insertion-Removal Inpainting. *arXiv preprint arXiv:2503.08677* (2025).
- Yuchen Zeng, Wonjun Kang, Yicong Chen, Hyung Il Koo, and Kangwook Lee. 2024. Can MLLMs Perform Text-to-Image In-Context Learning? *arXiv preprint arXiv:2402.01293* (2024).
- Yuxin Zhang, Minyan Luo, Weiming Dong, Xiao Yang, Haibin Huang, Chongyang Ma, Oliver Deussen, Tong-Yee Lee, and Changsheng Xu. 2025. IP-Prompter: Training-Free Theme-Specific Image Generation via Dynamic Visual Prompting. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. 1–12.
- Chenyang Zhu, Kai Li, Yue Ma, Longxiang Tang, Chengyu Fang, Chubin Chen, Qifeng Chen, and Xiu Li. 2024. InstantSwap: Fast Customized Concept Swapping across Sharp Shape Differences. In *The Thirteenth International Conference on Learning Representations*.