Inpaint-Anywhere: Zero-Shot Multi-Identity Inpainting with Efficient Diffusion Transformer

Junsheng Luan, Lei Zhao*, Wei Xing

College of Computer Science and Technology, Zhejiang University
{l.junsheng121, cszhl, wxing}@zju.edu.cn

Abstract

Subject-driven generation, which aims to synthesize visual content for a given identity V^* with specific attributes, has garnered increasing attention in recent years. While existing methods demonstrate impressive identity consistency for both single and multiple identities, they often lack user-specified spatial control. Recent approaches, such as OminiControl-2 and EasyControl, enable inpainting conditioned on a single identity but fall short in multi-identity scenarios. In this paper, we introduce **BoundID**, a dataset synthesis pipeline for generating multi-identity images with bounding box annotations, and introduce **Inpaint-Anywhere**, a diffusion transformer framework for multi-identity inpainting. Given multiple identity references and corresponding masks, our method simultaneously generates all desired identities at precise locations while achieving both high identity and prompt fidelity. Extensive experiments show that Inpaint-Anywhere achieves state-of-the-art performance in multi-identity inpainting.

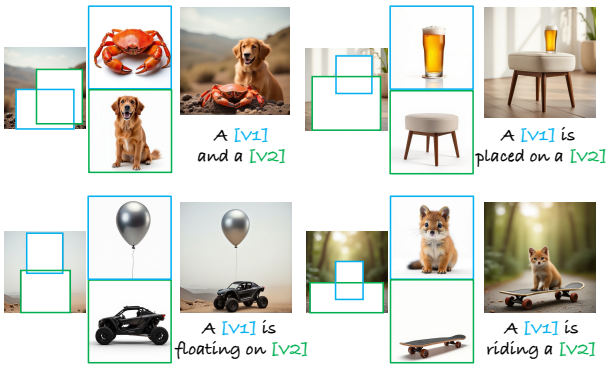

Figure 1: Given reference images of multiple V^* and corresponding spatial masks, our method generates all desired identities simultaneously at precise user-specified locations, while maintaining high identity and prompt fidelity.

Introduction

Subject-driven image generation (Ruiz et al. 2023; Kumari et al. 2023; Gal et al. 2023; Wei et al. 2023; Ye et al. 2023; Shi et al. 2024; Zhang et al. 2024a; He et al. 2024; Luan et al. 2025c; Li et al. 2025a,b; Luan et al. 2025a,b) aims to generate images for a given identity. The identity, denoted as V^*, represents an entity with distinctive visual attributes, such as a dog-shaped backpack. The core requirement in this task is to ensure both identity fidelity (i.e., preserving the visual attributes of V^*) and prompt fidelity (i.e., alignment with the prompt inputs) of the generated images. Existing approaches can be categorized into tuning-based methods and tuning-free methods. Tuning-based methods (Ruiz et al. 2023; Kumari et al. 2023) fine-tune a pre-trained diffusion model using the reference images of V^*. They achieve high identity fidelity but also bring additional computational cost. Tuning-free methods (Wei et al. 2023; Ye et al. 2023; Shi et al. 2024; Zhang et al. 2024a; Wu et al. 2025) directly extract features from the reference images of V^*, avoiding fine-tuning and thus gaining increasing attention from the community. Recently, multi-identity generation methods (Wu et al. 2025; He et al. 2025; Chen et al. 2024) synthesize multiple identities within a single scene. This paradigm extends the flexibility and applicability of subject-driven generation, enabling cases such as personalized comics and story-driven visual synthesis (Cao et al. 2025; Ma et al. 2025; Feng et al. 2024; Wang et al. 2025).

However, most of these approaches lack user-specified spatial controllability, which limits their capacity for fine-grained and precise control in subject-driven generation. To address this problem, methods such as AnyDoor (Chen et al. 2024), OminiControl-2 (Tan et al. 2025b) and EasyControl (Zhang et al. 2025b) have been proposed, allowing users to guide the subject-driven generation process via additional mask, **i.e., subject-driven inpainting**. Although they achieve promising results, they are limited to single-identity scenarios. Generating multiple identities with user-specified masks, **i.e., multi-identity inpainting**, requires multiple inference passes, which is time-consuming and often leads to visual artifacts and boundary inconsistencies.

Given references of multiple V^* and their corresponding spatial masks, our method performs high-quality multi-identity inpainting by generating all target identities at user-specified locations while maintaining high identity and prompt fidelity. To this end, we tackle two key challenges.

The first challenge lies in the lack of suitable multi-

*Corresponding author.

identity datasets with explicit spatial annotations. Existing datasets (Tan et al. 2025a; Wu et al. 2025) focus primarily on single-identity or multi-identity scenarios but lack spatial annotations. To bridge this gap, we introduce a data synthesis pipeline, **BoundID**, which generates multi-identity images annotated with bounding boxes. Specifically, we first construct identity pairs by sampling from the Subject200k dataset (Tan et al. 2025a), then use the multi-identity generator UNO (Wu et al. 2025) to synthesize the candidate multi-identity images. We then apply Grounding DINO (Liu et al. 2024) along with category labels for identity-level detection. To ensure accurate identity generation and annotation, we apply a three-stage filtering process: (1) retain samples where all identities are successfully detected and boxed; (2) verify consistency between reference identities and detected regions using GPT-4o; and (3) select aesthetically high-quality samples, also assessed by GPT-4o. We ultimately generate 43k high-quality multi-identity images with bounding boxes.

The second challenge lies in generating all target identities at precise, user-specified locations, which involves addressing two critical sub-problems: (1) efficiently fusing spatial masks with identity features, and (2) ensuring accurate identity-to-region correspondence. To address the first sub-problem, we propose **Inpaint-Anywhere**, which leverages the diffusion transformer (DiT) (Labs 2023) backbone to seamlessly integrate identity reference images with spatial masks. Specifically, we decompose the subject-driven inpainting task into the spatially aligned task (image inpainting) and the non-spatially aligned task (subject-driven generation), then introduce lightweight inpainting and identity adapters to handle each component. This approach enables efficient multi-modal attention interaction between the identity and mask features.

For the second sub-problem, the model struggles to associate each identity with its designated region when handling multiple identities and spatial masks. For instance, identity V_1^* may appear in the region intended for V_2^*. Therefore, we propose a **position encoder** that encodes the mask boxes into multiple position tokens. These tokens are then used to conduct **masked position-aware attention** within the diffusion transformer. Specifically, each target identity is assigned a spatial region defined by the mask, and the corresponding position tokens are injected into the model within the attention block. The position tokens act as an extra spatial context, with the noisy image tokens attending to them, then the model achieves identity-to-region correspondence.

In this manner, Inpaint-Anywhere performs high-quality multi-identity inpainting by generating all target identities at user-specified locations while maintaining high identity and prompt fidelity. It does not require fine-tuning of specific V^* and introduces only 48.7M additional parameters. Our contributions are summarized as follows:

- We introduce a data synthesis pipeline **BoundID**, which generates multi-identity images with bounding box annotations, combining multi-identity generation, identity-level detection and GPT-4o filtering, addressing the lack of suitable datasets for multi-identity inpainting.
- We propose **Inpaint-Anywhere**, a multi-identity inpainting model to generate all target identities at user-specified locations while maintaining high identity and prompt fidelity. It leverages DiT's backbone to integrate reference images with spatial masks and employs lightweight adapters for efficient multi-modal attention.
- We propose a **position encoder** and **masked position-aware attention**, enabling the model to achieve identity-to-region correspondence.
- Extensive experiments show that Inpaint-Anywhere achieves state-of-the-art performance in multi-identity inpainting.

Related Work

Subject-driven Generation

With the rapid development of diffusion models (Song et al. 2025a; Huang et al. 2025; Zhang et al. 2025a; Chen et al. 2025; Huang et al. 2024; Zhou et al. 2024; Song et al. 2025b; Cao et al. 2024; Zhang et al. 2024c, 2025d,c, 2024b), subject-driven generation has gained significant attention for generating high-quality images of specific identities V^*. Tuning-free subject-driven generation methods (Wei et al. 2023; Ye et al. 2023; Shi et al. 2024; Zhang et al. 2024a; He et al. 2025) use an encoder to extract identity features from the reference image of V^* and inject them into the denoising network to guide the generation process. These methods eliminate the need for model fine-tuning required in tuning-based approaches (Ruiz et al. 2023; Kumari et al. 2023; Gal et al. 2023).

Specifically, ELITE (Wei et al. 2023) performs both global and local mapping techniques to integrate the visual feature of V^* into subject-driven text-to-image generation. InstantBooth (Shi et al. 2024) learns the general concepts of input images by converting input images into text tokens using a learnable image encoder. SSR-Encoder (Zhang et al. 2024a) uses a token-to-patch aligner to highlight the selective regions in the reference image by the query. Recently, UNO (Wu et al. 2025) proposes a multi-identity subject-driven generation with cross-modal alignment and universal rotary position embedding, achieving high consistency while ensuring controllability in both single-subject and multi-subject driven generation.

Conditional Diffusion Generation

Complex image synthesis tasks like subject-driven inpainting require fine-grained control. To this end, recent researches add conditional image control to diffusion models with extra networks. For instance, ControlNet (Zhang and Agrawala 2023) and T2I-Adapter (Mou et al. 2023) propose to fine-tune an extra reference network that encodes spatial information such as edges, depth, and human pose, to control the diffusion model together with text prompts. IP-Adapter (Ye et al. 2023) proposes to use an extra image encoder to extract high-level semantic features of reference images, and control image generation with both textual and visual prompts. These methods bring extra network modules and substantial training parameters. Recently, OminiControl (Tan et al. 2025a) leverages DiT's backbone to integrate

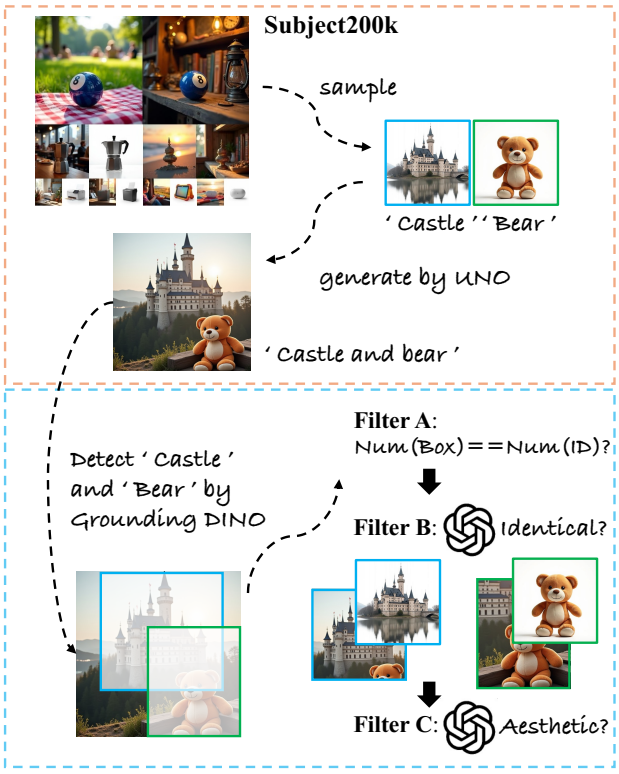

Figure 2: The BoundID pipeline. For the upper part, we construct identity pairs and synthesize the candidate multi-identity images. For the lower part, we generate bounding boxes and employ a three-stage filtering process.

conditional features with lightweight LoRA, achieving high-quality conditional image generation like image inpainting, subject-driven generation and sketch-to-image generation. Later, OminiControl-2 (Tan et al. 2025b) and EasyControl (Zhang et al. 2025b) are proposed to support multiple conditions. However, these methods are limited to single-identity inpainting. In this paper, we follow these methods to utilize the intrinsic backbone of diffusion transformer to achieve multi-identity inpainting.

Methods

We introduce the data synthesis pipeline **BoundID** and the multi-identity inpainting model **Inpaint-Anywhere**.

BoundID

UNO (Wu et al. 2025) introduces a high-resolution and identity-consistent data synthesis pipeline, enabling the generation of multi-identity datasets and achieving impressive performance in multi-identity generation tasks. However, this pipeline lacks explicit spatial annotations, which are crucial for training a spatially controllable generation model. Motivated by this limitation, we propose a data synthesis pipeline, denoted BoundID, which generates multi-identity images with bounding box annotations.

As shown in **Fig. 2**, for the upper part, we construct identity pairs by sampling from the Subject200k dataset (Tan et al. 2025a). Specifically, Subject200k contains numerous image pairs, each depicting the same identity under varying conditions, along with a corresponding category label. We randomly sample one image from each of two distinct identity pairs, forming a two-identity image pair. Then, we use the two-identity image pairs and leverage the multi-identity generation model UNO to synthesize the candidate multi-identity images.

For the lower part, we apply Grounding DINO (Liu et al. 2024) along with category labels for identity-level detection. Grounding DINO is a powerful open-set object detector that accurately localizes textual queries within images. Given a category label, it identifies the corresponding identity region in the generated multi-identity image and produces bounding boxes. In practice, we find that using an enlarged bounding box leads to better training performance, as it allows the identity to interact with more background context. This results in fewer inconsistencies around the mask boundaries during generation. To ensure accurate identity generation and annotation, we apply a three-stage filtering process: (1) retain samples where all identities are successfully detected and boxed; (2) verify consistency between reference identities and detected regions using GPT-4o; and (3) select aesthetically high-quality samples, also assessed by GPT-4o.

Our final dataset consists of 43k groups of $\{$multi-identity image X_{gt}, reference images of the identities $X_{1,2}$, identity-specific bounding boxes, corresponding masked multi-identity image X_m generated based on these boxes$\}$. Together, these components form a comprehensive multi-identity dataset with explicit spatial annotations.

Inpaint-Anywhere

Model Overview Our proposed **Inpaint-Anywhere** leverages the diffusion transformer (DiT) (Labs 2023) backbone to seamlessly integrate identity reference images with spatial masks. Specifically, we decompose the subject-driven inpainting task into spatially aligned task (image inpainting) and non-spatially aligned task (subject-driven generation), then introduce lightweight inpainting and identity adapters to handle each component.

To ensure accurate identity-to-region correspondence, we propose a **position encoder** and **masked position-aware attention**, enabling the model to achieve identity-to-region correspondence.

The training inputs are generated using the BoundID pipeline. For simplicity in illustration, we omit the textual condition.

Integrating Identity and Spatial Features To achieve subject-driven inpainting, during the training process, the ground truth X_{gt}, $X_{1,2}$, X_m are first encoded into latent space tokens using a frozen VAE. We then add Gaussian noise to X_{gt}, **i.e.** the noisy image tokens. These tokens are integrated into a unified sequence: $[X_{gt}; X_{1,2}; X_m]$, and are then normalized, scaled and shifted, then passed into the attention blocks (Peebles and Xie 2023). This enables direct interaction in multi-modal attention without additional reference networks or fine-tuning on the identities. The attention

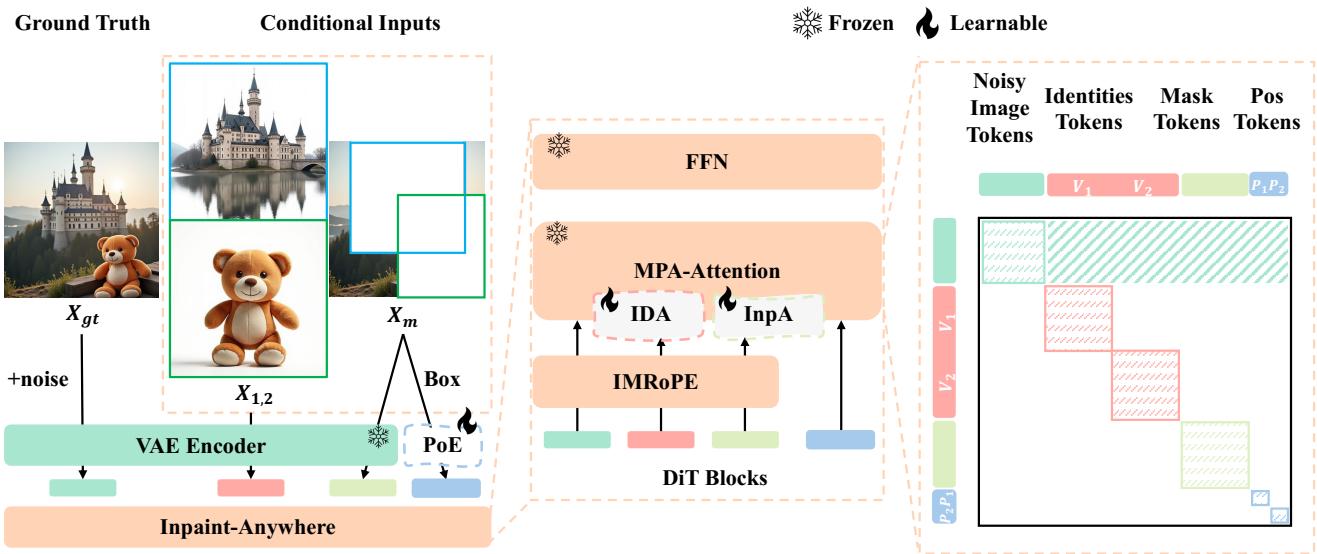

Figure 3: **Inpaint-Anywhere** leverages the diffusion transformer (DiT) backbone to seamlessly integrate identity reference images with spatial masks. Specifically, we decompose the subject-driven inpainting task into spatially aligned task (image inpainting) and non-spatially aligned task (subject-driven generation), then introduce lightweight **inpainting adapter (InpA)** and **identity adapter (IDA)** to handle each component. We propose a **position encoder (PoE)** and **masked position-aware attention**, enabling the model to achieve identity-to-region correspondence.

blocks process the subject-driven inpainting task by seamlessly integrating $X_{1,2}$ and X_m alongside X_{gt}, facilitating interaction between the identity and mask features.

Parameter-Efficient Joint Training For better controllability and fine-grained feature learning, we explicitly decouple the subject-driven inpainting task into the spatially aligned task (image inpainting) and the non-spatially aligned task (subject-driven generation), then introduce lightweight inpainting and identity adapters to handle each component (Luan et al. 2025b). We further propose the adapter switch to better control the adapters. An adapter has three statuses: *disabled*, *frozen*, and *learning*. The tokens will not be affected by a *disabled* adapter (by setting the adapter scale to zero); a *frozen* adapter affects the tokens while the parameters of the adapter are not optimized.

We add IDA, InpA, and the adapter switches to the QKV Linear layers in the attention blocks. For the task separation, $X_{1,2}$ is projected to QKV tokens with InpA *disabled*; X_m is projected with IDA *disabled*; X_{gt} is projected with both adapters *disabled*, indicating the use of the vanilla FLUX.1 image generation capability. The parameter-efficient joint training brings only 26.5M additional parameters from IDA and 13.2M additional parameters from InpA.

IMRoPE FLUX.1 proposes the Rotational Position Embedding (RoPE) to incorporate positional dependencies across tokens. We also perform it on $X_{1,2}$ and X_m to ensure effective interaction with the noisy image tokens X_{gt}. For a 512×512 identity/mask image, the VAE encodes it and divides it into 32×32 grid of tokens, then each token is assigned a unique two-dimensional position index (j, i) with $i \in [0, 32)$ and $j \in [0, 32)$. Following (Tan et al. 2025a), we propose the Identity-Mask Rotational Position Embedding, denoted IMRoPE, to ensure effective interaction among the tokens. Specifically, for the spatially aligned task (image inpainting), we keep the position index to the original settings in FLUX.1; for the non-spatially aligned task (subject-driven generation), we shift the position index j to $[32, 64)$ to avoid spatial overlap. This leads to faster convergence during training. After the training process, the model can generate the identities simultaneously at masked regions.

Position Encoder Currently, the model struggles to associate each identity with its designated region when handling multiple identities and spatial masks. For instance, identity V_1^* may be incorrectly generated in the region intended for V_2^*. To address this issue, we leverage the bounding boxes provided in the input as spatial priors and feed them into an extra position encoder to obtain position tokens that are uniquely assigned to each identity. Specifically, each bounding box is represented by its normalized top-left and bottom-right coordinates, forming a 4-dimensional vector. We transform them using Fourier feature encoding, then project them through a learnable multi-layer perceptron (MLP) to obtain a 3072-dimensional token $P_{1,2}$, same as the dimension of the noisy image tokens. This introduces only 9M additional parameters. By implicitly learning front-back and occlusion relationships from the training data, the model ensures that V_i^* is bound to P_i, making sure they are generated in their specified positions, and when the mask regions overlap, the model still produces visually consistent compositions.

Masked Position-Aware Attention Finally, we propose the Masked Position-Aware Attention mechanism. Each position token obtained from the encoded bounding box is

treated as an independent token and directly injected into the attention module alongside the noisy image tokens, identity tokens, and mask tokens. The position tokens act as an extra spatial context, with the noisy image tokens attending to them. This allows the model to ensure that both identities are generated in their corresponding locations, achieving identity-to-region correspondence.

Existing subject-driven diffusion models face an efficiency bottleneck due to the repeated computation of reference features at each denoising step. While the noisy image representation changes progressively as noise is removed across timesteps, the condition inputs remain unchanged throughout the sampling process (Tan et al. 2025b). To improve efficiency, as illustrated in the right of **Fig. 3**, we adopt the approach from OminiControl-2 (Tan et al. 2025b) by leveraging cacheable features and asymmetric attention masking. An attention mask is applied to prevent identity, mask, and position tokens from attending to noisy tokens, restricting them to attending only to themselves. This asymmetric setup, applied during training, ensures that noisy tokens benefit from conditioning information, while the identity, mask, and position tokens remain read-only and independent of the evolving noisy image. During inference, the key-value projections of these tokens are computed just once at the first denoising step, cached, and reused throughout, significantly reducing computational overhead.

The training uses the following denoising loss function:

$$L = \mathbb{E}_{\epsilon,t\sim\mathcal{U}(t)}\left[w(t)\left\|\epsilon_\theta(X_t,[X_{1,2};X_m;P_{1,2}]) - \epsilon\right\|^2\right] \tag{1}$$

where $X_t = (1-t)X_{gt} + t\epsilon$ is the noisy image tokens, ϵ_θ is the denoising DiT with IDA, InpA and PoE parameters θ, and $w(t)$ is a weighting function at timestep t, $\epsilon \sim N(0, I)$.

Experiments

Experimental Setup

Datasets From the Subject200k dataset, we sample 50k identities. We use the UNO framework to generate 100k two-identity images. After the three-stage filtering process, we ultimately retain 43k groups of {multi-identity image X_{gt}, reference images of the identities $X_{1,2}$, identity-specific bounding boxes, corresponding masked multi-identity image X_m generated based on these boxes}. We randomly sample 2k groups for validation (using $X_{1,2}$, X_m as inputs) while the remaining groups are used for training. We generate 4 images for each group, totally 8,000 images. We further validate our approach using the DreamBench dataset (Ruiz et al. 2023), which contains a total of 30 identities. We randomly create 150 two-identity image pairs and use GPT-4 to generate four validation prompts for each pair, containing placeholders for the identities ("[V1]" and "[V2]"), a background prompt, and two sets of bounding box coordinates. The background prompt is fed into a pretrained FLUX model to generate a background image. Binary masks are then applied based on the provided bounding boxes. Each validation prompt results in the generation of two multi-identity images conditioned on the masked background and the respective two-identity image pair, yielding 8 images per pair, totally 1,200 images.

Implementation Details Our base model is FLUX.1-dev. We implement the adapters with LoRA (Hu et al. 2021). The rank of IDA to 8, while the rank of InpA is set to 4. The model is trained with a batch size of 6 and an accumulation step of 1. We employ the Prodigy optimizer (Mishchenko and Defazio 2024) with safeguard warmup and bias correction enabled, setting the weight decay to 0.01. We conduct training on four NVIDIA A100 (80GB) GPUs with image resolutions of 512×512. The inference step is 25. We use "[V1]", "[V2]" as identity placetokens.

Baselines We collect three methods that support single-identity image inpainting, including IP-Adapter (Ye et al. 2023) (with Controlnet) (Zhang, Rao, and Agrawala 2023), AnyDoor (Chen et al. 2024), and EasyControl (Zhang et al. 2025b). By performing multiple inference passes, these methods can achieve multi-identity image inpainting. All images are generated at a resolution of 512×512.

Evaluation Metrics We evaluate these methods on (1) *CLIP-I*, (2) *DINO*, (3) *CLIP-T*, (4) *FID* and (5) *human preference*. For (1) *CLIP-I*, we compute the identity fidelity for each individual identity. Specifically, for each identity V^*, we extract the corresponding box region from the generated image and its reference image. These images are then encoded into embeddings using a pretrained ViT-L/16 model. We report average cosine similarity between the embeddings, which evaluates the preservation of identity details in the generated images. For (2) *DINO* (Zhang et al. 2022), we calculate cosine similarity between the ViT-S/16 DINO embeddings. This encourages the distinction of unique features of an identity or image, which is considered better than CLIP-I (Ruiz et al. 2023). For (3) *CLIP-T*, we compute CLIP embeddings of the generated image and the prompt (replacing the placeholder with corresponding category) and then calculate the CLIP distances of the CLIP embeddings. *CLIP-I* and *DINO* indicate the ID fidelity, and *CLIP-T* indicates the prompt fidelity. For (4) *FID*, we employ the Frechet Inception Distance (FID) to evaluate the realism of the generated results. For (5) *human preference*, we collect 200 online questionnaires. Each questionnaire shows four images generated from the four methods. The questionnaire then proposes questions A & B: "A. Which of the following four images preserves the most details of the V^*? " (evaluation on identity fidelity), "B. Which of the following best corresponds to the prompts?" (evaluation on prompt fidelity).

Qualitative Results

For the single-identity inpainting methods, IP-Adapter (with ControlNet), AnyDoor, and EasyControl, we perform multiple inference passes to handle multi-identity image inpainting. As shown in **Fig. 4**, IP-Adapter (with ControlNet) and AnyDoor fail to preserve identity features effectively and introduce visual artifacts and inconsistencies near mask boundaries. EasyControl performs better in preserving identity features and mitigating boundary artifacts, but still struggles with capturing fine-grained details such as color and shape. In contrast, Inpaint-Anywhere achieves state-of-

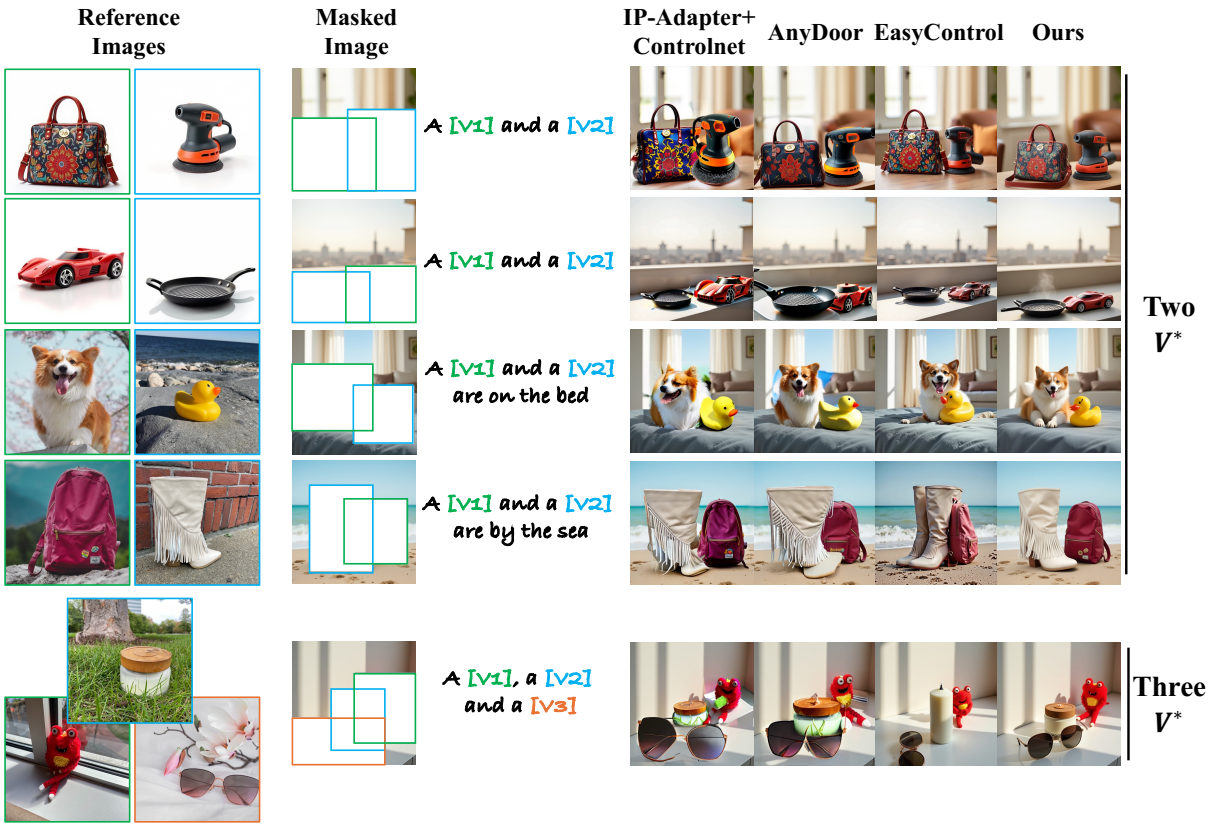

Figure 4: The qualitative comparison. For the three single-identity image inpainting, we perform multiple inference passes to conduct multi-identity image inpainting. In contrast, Inpaint-Anywhere achieves state-of-the-art performance in multi-identity inpainting, while requiring only a single inference pass.

the-art performance in multi-identity image generation with precise spatial controllability, while requiring only a single inference pass.

Quantitative Results

We conduct a comprehensive quantitative comparison across the DreamBench and BoundID datasets, as well as a human preference study. As shown in **Tab. 1**, our method outperforms existing approaches across all evaluation metrics, while requiring only a single inference pass and 6.53 seconds for each image.

Ablation Study

For the **BoundID** pipeline, we conduct ablation study on (1) BoundID filters, (2) enlarged bounding box; for **Inpaint-Anywhere**, we conduct ablation study on (3) IDA and InpA, (4) IMRoPE, and (5) position encoder. We conduct the experiments on dataset generated by BoundID.

BoundID Filters We further evaluate how each filtering stage impacts the overall dataset quality and the model performance in **Tab. 2**.

Enlarged Bounding Box When applying Grounding DINO to generate bounding boxes, we find that using enlarged boxes leads to better training performance, as they

Methods	IP-Adapter +Controlnet	AnyDoor	EasyControl	**Ours**
DreamBench				
CLIP-I↑	0.733	0.815	0.803	**0.881**
DINO↑	0.601	0.685	0.727	**0.745**
CLIP-T↑	0.261	0.281	0.286	**0.305**
FID↓	12.21	10.84	8.761	**7.619**
BoundID				
CLIP-I↑	0.686	0.769	0.761	**0.823**
DINO↑	0.549	0.621	0.700	**0.721**
CLIP-T↑	0.237	0.255	0.269	**0.285**
FID↓	14.73	11.31	9.295	**8.521**
Human Preference				
Identity Fidelity	0.01	0.13	0.25	**0.61**
Prompt Fidelity	0.03	0.22	0.26	**0.49**
Efficiency				
Inference Passes	2	2	2	**1**
Inference Time (s)	10.7	12.8	15.7	**6.53**

Table 1: Quantitative comparison. The bold data indicates the best. Our method outperforms existing approaches across all evaluation metrics, while requiring only a single inference pass and 6.53 seconds each image.

	CLIP-I↑	DINO↑	CLIP-T↑	FID↓	Image Num
Only A	0.616	0.669	0.267	9.365	72k
A & B	0.741	0.701	0.256	16.25	51k
A & B & C (Ours)	**0.823**	**0.721**	**0.285**	**8.521**	**43k**

Table 2: Ablation study on BoundID filters. The Filters A,B,C improves the overall dataset quality and the model performance.

Box Edge	CLIP-I↑	DINO↑	CLIP-T↑	FID↓
Original	0.811	**0.723**	0.280	10.04
+25%	0.813	0.719	0.276	9.256
+50% (Ours)	**0.823**	0.721	**0.285**	**8.521**
+75%	0.820	0.703	0.261	8.678

Table 3: Ablation study on enlarged bounding boxes. A bounding box extended by 50% along each edge achieves the best performance.

	CLIP-I↑	DINO↑	CLIP-T↑	FID↓
w/o IDA	0.616	0.669	0.267	9.362
w/o InpA	0.741	0.701	0.256	16.25
Unified Adapter	0.795	0.683	0.278	11.57
Ours	**0.823**	**0.721**	**0.285**	**8.521**

Table 4: Ablation study on IDA and InpA. Results show that the proposed IDA and InpA are pivotal for the multi-identity inpainting task; the separated adapters have better controllability and fine-grained feature learning.

allow the identity to interact with more background context. As shown in **Fig. 3**, a bounding box extended by 50% along each edge (clipped within image) yields the best performance. FID score improves compared to the original box, due to enhanced consistency around the mask boundaries.

IDA and InpA To evaluate the impact of explicitly decomposing the subject-driven inpainting task into spatially aligned task (image inpainting) and non-spatially aligned task (subject-driven generation), we disable the IDA or InpA to evaluate their effect on the subject-driven inpainting task. As shown in **Tab. 4**, without IDA, the identity feature cannot be integrated into the subject-driven inpainting task, which decreases the CLIP-I value; without InpA, the identity is not filled in the correct position, bringing artifacts and decreasing the visual fidelity.

We also train a unified adapter instead of the separated adapters to evaluate the effect of the decomposed adapter design. The last two rows of **Tab. 4** indicates better controllability and fine-grained feature learning of separated adapters.

IMRoPE As shown in **Tab. 5**, due to the RoPE spatial overlap, there is a significant drop in both DINO and CLIP-I scores when cloning the position index from the target image without using IMRoPE.

Position Encoder As shown in **Fig. 5**, without position encoder, the identities are sometimes not generated in their

	CLIP-I↑	DINO↑	CLIP-T↑	FID↓
w/o IMRoPE	0.725	0.607	0.273	11.31
Ours	**0.823**	**0.721**	**0.285**	**8.521**

Table 5: Ablation study on IMRoPE. IMRoPE improves the overall performance by avoiding RoPE spatial overlap.

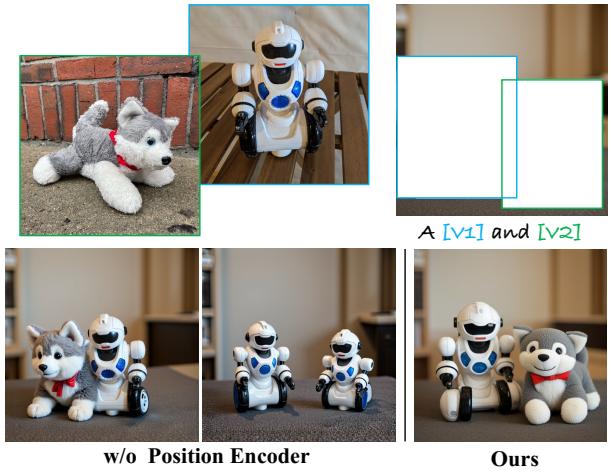

Figure 5: Without position encoder, the identities are sometimes not generated in their designated masked regions.

designated masked regions: identity V_1^* is generated in the mask designated for V_2^*, or identity V_1^* is generated in all the masks.

Limitation

Our method exhibits several limitations primarily due to the constraints of the constructed training dataset. First, the inpainting masks are required to be relatively regular in shape (square) and size (covering a large portion of the given image), since the proposed model is trained under such conditions. Second, since the training dataset does not include facial images, our method may fail to synthesize high-fidelity facial images. Lastly, our current framework supports generation at a fixed resolution of 512×512, which restricts its applicability to higher-resolution tasks.

Conclusion

We present Inpaint-Anywhere, a novel diffusion transformer framework for multi-identity image inpainting. By applying the BoundID dataset synthesis pipeline, Inpaint-Anywhere performs high-quality multi-identity inpainting by generating all target identities at user-specified locations while maintaining high identity and prompt fidelity. Experimental results demonstrate that our method achieves state-of-the-art performance in multi-identity inpainting. For future work, we aim to enrich the dataset with a broader range of mask shapes and sizes and facial identities, extend support to higher-resolution synthesis, and enhance the generalization capability of the model on real-world photographic images.

Acknowledgments

This work was supported in part by Zhejiang Province Program (2024C03263, 2022C01222, 2023C03199, 2023C03201, LZ25F020006), the National Program of China (62172365, 19ZDA197), Macau project: Key technology research and display system development for new personalized controllable dressing dynamic display, Ningbo Science and Technology Plan Project (2025Z052, 2025Z062, 2022Z167, 2023Z137), and MOE Frontier Science Center for Brain Science & Brain-Machine Integration (Zhejiang University).

References

Cao, K.; He, X.; Hu, T.; Xie, C.; Zhang, J.; Zhou, M.; and Hong, D. 2024. Shuffle mamba: State space models with random shuffle for multi-modal image fusion. *arXiv preprint arXiv:2409.01728*.

Cao, K.; Wang, J.; Ma, A.; Feng, J.; Zhang, Z.; He, X.; Liu, S.; Cheng, B.; Leng, D.; Yin, Y.; et al. 2025. Relactrl: Relevance-guided efficient control for diffusion transformers. *arXiv preprint arXiv:2502.14377*.

Chen, S.; Bai, J.; Zhao, Z.; Ye, T.; Shi, Q.; Zhou, D.; Chai, W.; Lin, X.; Wu, J.; Tang, C.; et al. 2025. An empirical study of gpt-4o image generation capabilities. *arXiv preprint arXiv:2504.05979*.

Chen, X.; Huang, L.; Liu, Y.; Shen, Y.; Zhao, D.; and Zhao, H. 2024. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6593–6602.

Feng, J.; Ma, A.; Wang, J.; Cao, K.; and Zhang, Z. 2024. Fancyvideo: Towards dynamic and consistent video generation via cross-frame textual guidance. *arXiv preprint arXiv:2408.08189*.

Gal, R.; Arar, M.; Atzmon, Y.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2023. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)*, 42(4): 1–13.

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. *arXiv preprint arXiv:2501.09503*.

He, X.; Liu, Q.; Qian, S.; Wang, X.; Hu, T.; Cao, K.; Yan, K.; and Zhang, J. 2024. Id-animator: Zero-shot identity-preserving human video generation. *arXiv preprint arXiv:2404.15275*.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Huang, J.; Huang, Y.; Liu, J.; Zhou, D.; Liu, Y.; and Chen, S. 2025. Dual-Schedule Inversion: Training-and Tuning-Free Inversion for Real Image Editing. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 660–669. IEEE.

Huang, J.; Zhou, D.; Liu, J.; Shi, L.; and Chen, S. 2024. Ifast: Weakly supervised interpretable face anti-spoofing from single-shot binocular nir images. *IEEE Transactions on Information Forensics and Security*.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1931–1941.

Labs, B. F. 2023. FLUX. https://github.com/black-forest-labs/flux. Accessed: 2025-08-02.

Li, G.; Wang, Y.; Luan, J.; Zhao, L.; Xing, W.; Lin, H.; and Ou, B. 2025a. Cascaded diffusion models for virtual try-on: Improving control and resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 4689–4697.

Li, G.; Zheng, S.; Zhang, H.; Chen, J.; Luan, J.; Ou, B.; Zhao, L.; Li, B.; and Jiang, P.-T. 2025b. MagicTryOn: Harnessing Diffusion Transformer for Garment-Preserving Video Virtual Try-on. *arXiv preprint arXiv:2505.21325*.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, 38–55. Springer.

Luan, J.; Li, G.; Zhang, Z.; Zhao, L.; and Xing, W. 2025a. IP-Controller: Decomposition and Optimization of Cross-Attention Maps for Accurate Subject-Driven Text-to-Image Diffusion Generation. *IEEE Transactions on Circuits and Systems for Video Technology*. Early Access.

Luan, J.; Li, G.; Zhao, L.; and Xing, W. 2025b. Mc-vton: Minimal control virtual try-on diffusion transformer. *arXiv preprint arXiv:2501.03630*.

Luan, J.; Zhang, Z.; Xing, W.; and Zhao, L. 2025c. Personalized text-to-image generation with Large Language and Vision Assistant enhanced training. *Engineering Applications of Artificial Intelligence*, 161: 112116.

Ma, A.; Feng, J.; Cao, K.; Wang, J.; Wang, Y.; Zhang, Q.; and Zhang, Z. 2025. Lay2Story: Extending Diffusion Transformers for Layout-Togglable Story Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16102–16111.

Mishchenko, K.; and Defazio, A. 2024. Prodigy: An Expeditiously Adaptive Parameter-Free Learner. In *Forty-first International Conference on Machine Learning*.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; Shan, Y.; and Qie, X. 2023. T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models. arXiv:2302.08453.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4205.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

Shi, J.; Xiong, W.; Lin, Z.; and Jung, H. J. 2024. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8543–8552.

Song, Q.; Wang, X.; Zhou, D.; Lin, J.; Chen, C.; Ma, Y.; and Li, X. 2025a. HERO: Hierarchical Extrapolation and Refresh for Efficient World Models. *arXiv preprint arXiv:2508.17588*.

Song, Q.; Zhou, D.; Lin, J.; Shen, F.; Wang, J.; Hu, X.; Chen, C.; and Heng, P.-A. 2025b. SceneDecorator: Towards Scene-Oriented Story Generation with Scene Planning and Scene Consistency. *arXiv preprint arXiv:2510.22994*.

Tan, Z.; Liu, S.; Yang, X.; Xue, Q.; and Wang, X. 2025a. Ominicontrol: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14940–14950.

Tan, Z.; Xue, Q.; Yang, X.; Liu, S.; and Wang, X. 2025b. Ominicontrol2: Efficient conditioning for diffusion transformers. *arXiv preprint arXiv:2503.08280*.

Wang, J.; Ma, A.; Cao, K.; Zheng, J.; Zhang, Z.; Feng, J.; Liu, S.; Ma, Y.; Cheng, B.; Leng, D.; et al. 2025. Wisa: World simulator assistant for physics-aware text-to-video generation. *arXiv preprint arXiv:2503.08153*.

Wei, Y.; Zhang, Y.; Ji, Z.; Bai, J.; Zhang, L.; and Zuo, W. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15943–15953.

Wu, S.; Huang, M.; Wu, W.; Cheng, Y.; Ding, F.; and He, Q. 2025. Less-to-more generalization: Unlocking more controllability by in-context generation. *arXiv preprint arXiv:2504.02160*.

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. *arXiv preprint arXiv:2308.06721*.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. arXiv:2203.03605.

Zhang, L.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhang, S.; Xie, B.; Yan, Z.; Zhang, Y.; Zhou, D.; Chen, X.; Qiu, S.; Liu, J.; Xie, G.; and Lu, Z. 2025a. Trade-offs in Image Generation: How Do Different Dimensions Interact? *arXiv preprint arXiv:2507.22100*.

Zhang, Y.; Song, Y.; Liu, J.; Wang, R.; Yu, J.; Tang, H.; Li, H.; Tang, X.; Hu, Y.; Pan, H.; et al. 2024a. Ssr-encoder: Encoding selective subject representation for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8069–8078.

Zhang, Y.; Yuan, Y.; Song, Y.; Wang, H.; and Liu, J. 2025b. Easycontrol: Adding efficient and flexible control for diffusion transformer. *arXiv preprint arXiv:2503.07027*.

Zhang, Z.; Zhang, Q.; Lin, H.; Xing, W.; Mo, J.; Huang, S.; Xie, J.; Li, G.; Luan, J.; Zhao, L.; et al. 2024b. Towards highly realistic artistic style transfer via stable diffusion with step-aware and layer-aware prompt. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 7814–7822.

Zhang, Z.; Zhang, Q.; Luan, J.; Yang, M.; Wang, Y.; and Zhao, L. 2025c. SPAST: Arbitrary style transfer with style priors via pre-trained large-scale model. *Neural Networks*, 107556.

Zhang, Z.; Zhang, Q.; Luan, J.; Yang, M.; Wang, Y.; and Zhao, L. 2025d. VectorSketcher: Learning to create a vector-based free-hand sketch. *Engineering Applications of Artificial Intelligence*, 156: 111005.

Zhang, Z.; Zhang, Q.; Xing, W.; Li, G.; Zhao, L.; Sun, J.; Lan, Z.; Luan, J.; Huang, Y.; and Lin, H. 2024c. Artbank: Artistic style transfer with pre-trained diffusion model and implicit style prompt bank. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 7396–7404.

Zhou, D.; Huang, J.; Bai, J.; Wang, J.; Chen, H.; Chen, G.; Hu, X.; and Heng, P.-A. 2024. Magictailor: Component-controllable personalization in text-to-image diffusion models. *arXiv preprint arXiv:2410.13370*.