Insert Anything: Image Insertion via In-Context Editing in DiT

Wensong Song[1], Hong Jiang[1], Zongxing Yang[2], Zheqiao Cheng[1], Ruijie Quan[3], Yi Yang[1]*

[1]Zhejiang University
[2]Harvard University
[3]Nanyang Technological University
wensong.song@zju.edu.cn

Abstract

This work presents **Insert Anything**, a unified framework for reference-based image insertion that seamlessly integrates objects from reference images into target scenes under flexible, user-specified control guidance. Instead of training separate models for individual tasks, our approach is trained once on our new **AnyInsertion** dataset, the first open-source large-scale dataset specifically designed for reference image–based image editing, comprising 136K prompt-image pairs covering diverse tasks such as person, object, and garment insertion–and effortlessly generalizes to a wide range of insertion scenarios. Such a challenging setting requires capturing both identity features and fine-grained details, while allowing versatile local adaptations in style, color, and texture. To this end, we propose to leverage the multimodal attention of the Diffusion Transformer (DiT) to support both mask- and text-guided editing. Furthermore, we introduce an in-context editing mechanism that treats the reference image as contextual information, employing two prompting strategies to harmonize the inserted elements with the target scene while faithfully preserving their distinctive features. Extensive experiments on AnyInsertion, DreamBooth, and VTON-HD benchmarks demonstrate that our method consistently outperforms existing alternatives, underscoring its great potential in real-world applications such as creative content generation, virtual try-on, and scene composition.

Introduction

Recent advances in diffusion models (Ho, Jain, and Abbeel 2020; Peebles and Xie 2023; Yang, Zhuang, and Pan 2021) have revolutionized image editing (Brooks, Holynski, and Efros 2023; Zhang et al. 2023; Tan et al. 2024; Li et al. 2024; Zhou et al. 2024; Xu, Yang, and Yang 2025). Among various editing techniques, reference image-based editing plays a pivotal role by providing explicit visual cues that guide the editing process. In this approach–also referred to as image insertion, a target image is modified by seamlessly incorporating elements from a reference image, ensuring that the overall coherence and quality remain intact. Unlike text-based instructions, reference images offer concrete details regarding style, color, and texture, making them indispensable for achieving contextually consistent results.

*Corresponding author.

Despite promising advances that have been made, several challenges remain in current image insertion work. (1) **Task-Specific Focus.** Both approaches and datasets of them aim to address only specific tasks, such as either person insertion (Kulal et al. 2023) or garment insertion (Chong et al. 2024), which limits their broader applicability to real-world scenarios. (2) **Fixed Control Mode.** These methods typically rely on inflexible control mode, e.g., supporting either mask-guided editing (Chen et al. 2024b,a) that uses manually provided masks to delineate the editing region, or text-guided editing (He et al. 2024; Shen et al. 2025; Xu et al. 2024b) that depends on language instructions to control editing. This rigidity in control modes hinders creative flexibility. (3) **Inconsistent Visual-Reference Harmony.** Even when these methods (Chen et al. 2024b; He et al. 2024; Mao et al. 2025) succeed in inserting new elements, they often struggle to maintain visual harmony between the inserted content and the target image, while ensuring that the distinctive characteristics of the reference are retained. As a result, outputs usually exhibit artifacts or stylistic mismatches that compromise overall quality and authenticity.

First, to address the challenge (1), we introduce ***AnyInsertion***, the first open-source dataset for **multi-task, multi-modal** reference-based image editing. It is characterized by two key features: 1) **Diverse Task and Insertion Type Coverage.** It covers person, object, and garment insertion, with both *additive insertion* (e.g., adding a plant) and *replacement insertion* (e.g., swapping a chair). Categories include humans, daily items, garments, and furniture. This diversity ensures broader applicability to real-world scenarios. 2) **Multi-Modal Control Support.** It contains 136K prompt-image pairs, consisting of 58K mask-prompt pairs and 78K text-prompt pairs. These diverse control modes enable the dataset to provide flexible training and evaluation for both mask-guided and text-guided image editing tasks.

Building on our dataset, we introduce **Insert Anything**, a unified framework that inserts *any* reference element into a target scene via **multiple control modes** (i.e., mask and text). To address the challenge (2), we introduce ***multimodal attention control***, leveraging the multimodal attention of the Diffusion Transformer (DiT) (Peebles and Xie 2023) to jointly model the relationships among text, masks, and image patches, enabling flexible control over image editing tasks guided by either masks or text. Moreover, to address

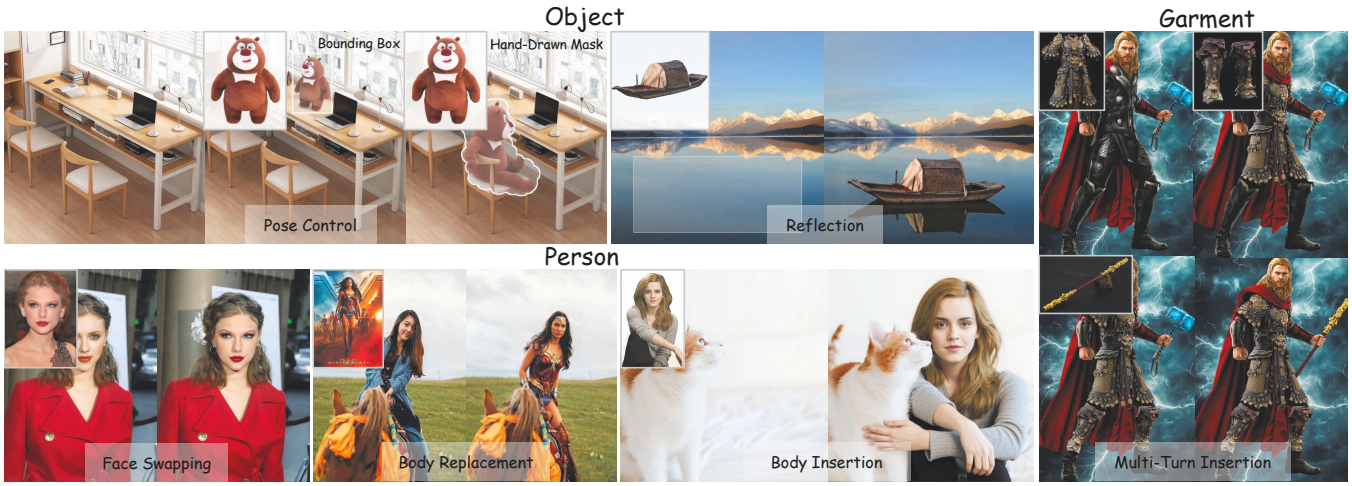


Figure 1: Our method supports diverse challenging scenarios across object, person, and garment insertion: object insertion with pose control (bounding box and hand-drawn mask) and reflection handling; person insertion with fine-grained edits (face swapping) and full-body manipulations (body replacement, insertion); garment insertion with multi-turn insertion.

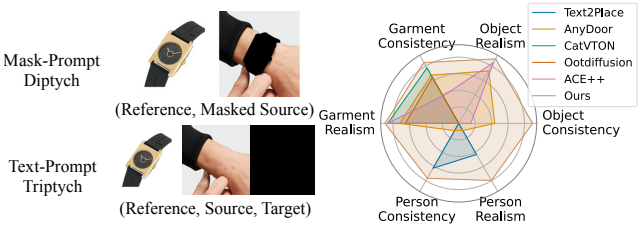


Figure 2: Prompt strategies (left) and performance (right). Methods missing specific task metrics (e.g., Ootdiffusion for object tasks) are assigned 0, indicating unsupported tasks.

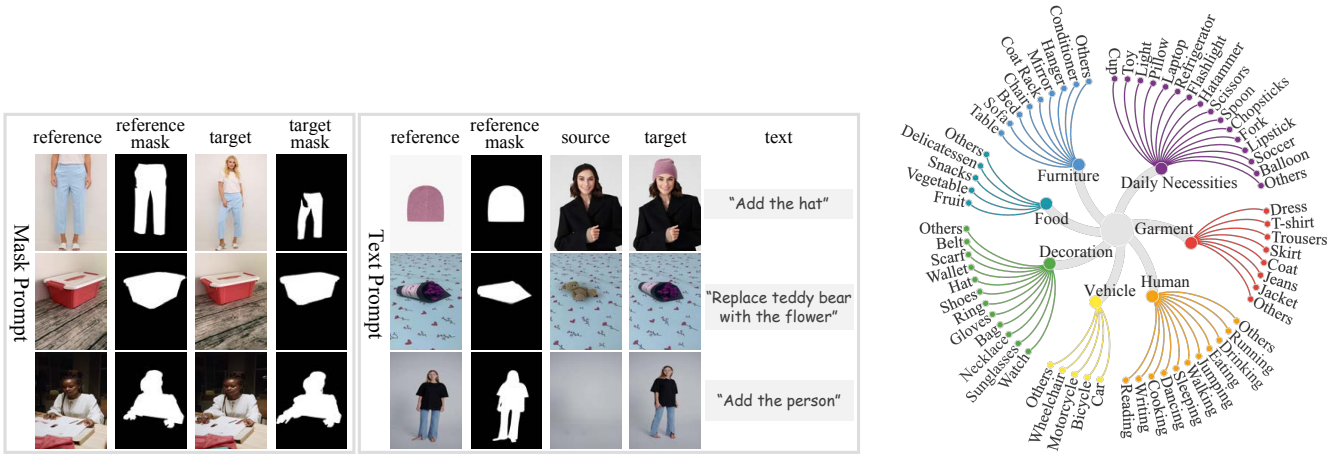
the challenge (3), we introduce ***in-context editing***, a new approach that treats the reference image as contextual content rather than as a standalone input. It allows for interactions between the inserted elements and their surrounding context, such as scale adjustment, consistent lighting and shadows, and viewpoint alignment, enabling the model to capture inherent correlations implicitly. To accommodate distinct control modes, we design two prompting strategies as in Fig. 2.

We perform extensive evaluations on our proposed ***AnyInsertion*** dataset and on two additional benchmarks–DreamBooth (Ruiz et al. 2023), and VTON-HD (Choi et al. 2021). As in Fig. 2, experimental results on human, object, and garment insertion tasks across these datasets demonstrate that our method achieves state-of-the-art performance. In summary, our **contributions** are three-fold:

- We introduce ***AnyInsertion***, a large-scale dataset containing 136K prompt-image pairs, spanning person, object, and garment insertion.
- We propose **Insert Anything**, a unified framework that seamlessly handles multiple insertion tasks (person, object, and garment) through a single model.
- We are the first, to our best knowledge, to leverage the DiT for image insertion, exploiting its unique capabilities for different control modes.
- We develop *in-context editing*, employing `diptych` and `triptych prompting` to integrate reference elements into target scenes while preserving identity.

Related Work

Image Insertion. Image insertion methods are usually categorized by task specificity and control strategy. Task-specific approaches include person insertion and garment insertion.
i) Person Insertion. Kulal et al. (Kulal et al. 2023) introduced an inpainting-based method, while ESP (Ostrek et al. 2024) generates personalized figures guided by 2D pose and scene context. Text2Place (Parihar et al. 2024) leverages SDS loss to optimize semantic masks for human placement.
ii) Garment Insertion. OOTDiffusion (Xu et al. 2024a) employs a ReferenceNet structure similar to a denoising UNet for processing garment images, whereas CatVTON (Chong et al. 2024) spatially concatenates garment and person images to enable lightweight virtual try-on.
iii) Object Insertion. Paint-by-Example (Yang et al. 2023), ObjectStitch (Song et al. 2022), AnyDoor (Chen et al. 2024b), Imprint (Song et al. 2024) and MimicBrush (Chen et al. 2024a) both support mask-guided insertion. Paint-by-Example and ObjectStitch utilize CLIP (Radford et al. 2021) image encoder to convert images into embedding, leveraging pretrained models for guidance. AnyDoor utilizes DINOv2 (Oquab et al. 2023) for feature extraction and ControlNet (Zhang, Rao, and Agrawala 2023) to preserve high-frequency details. Imprint learns view-invariant identity-preserving representations via multi-view object pairs before training composition models. MimicBrush uses a UNet (Ronneberger, Fischer, and Brox 2015) to extract reference features while maintaining scene context via depth maps and unmasked background latents. Freeedit (He et al. 2024) supports text-guided insertion through multi-modal



(a) Data example. The dataset is divided into mask-prompt and text-prompt categories, with further subdivisions into accessories, objects, and persons for each prompt type.

(b) AnyInsertion covers diverse scenarios: furniture, apparel, vehicles, etc.

Figure 3: Overview of AnyInsertion dataset, highlighting its example (a) and diversity (b).

instructions with UNet-based fine-detail extraction.

Our approach differs from these methods by leveraging in-context learning for efficient high-frequency detail extraction, eliminating the need for additional networks like ControlNet, and supporting both mask and text prompts.

Unified Image Generation and Editing. Recent frameworks have attempted to unify multiple image generation and editing tasks. OmniGen (Xiao et al. 2025) tokenizes text and images into a long tensor, jointly modeling them within a single model to achieve unified representations across modalities. ACE (Han et al. 2024) employs a conditioning unit for multiple inputs, OminiControl (Tan et al. 2024) concatenates condition tokens with image tokens. AnyEdit (Yu et al. 2024), Unireal (Chen et al. 2024c), and Ace++ (Mao et al. 2025) provide partial support for image insertion tasks, but none offers a comprehensive solution for all three insertion types with both mask and text prompt support, which distinguishes our Insert Anything framework.

AnyInsertion Dataset

To enable diverse image insertion tasks, we introduce AnyInsertion, the first open-source, large-scale dataset for reference-based image editing. We first compare AnyInsertion with prior datasets, then describe our dataset construction process, and finally present detailed statistics.

Comparison with Existing Datasets

Existing datasets suffer from several limitations: (1) **Limited Data Categories.** FreeEdit (He et al. 2024) dataset primarily focuses on animals and plants, and the VITON-HD (Choi et al. 2021) dataset specializes in garments. Even AnyDoor (Chen et al. 2024b) and MimicBrush (Chen et al. 2024a) include a large scale of data, they contain only very few samples related to person insertion. (2) **Restricted Prompt Types.** FreeEdit provides only text-prompt data, while VITON-HD supports only mask-prompt data. (3) **Insufficient Image Quality.** AnyDoor and MimicBrush utilize a large volume of video data. These video datasets often suffer from low resolution and motion blur. To address these issues, we have constructed an AnyInsertion dataset.

Dataset	Theme	Resolution	Prompt	#Edits
FreeBench	Daily Object	256 × 256	Text	131,160
VITON-HD	Garment	1024 × 768	Mask	11,647
AnyInsertion	Multifield	Mainly 1–2 K	Mask / Text	159,908

Table 1: Comparison of existing image insertion datasets with AnyInsertion, which covers diverse object categories, supports both mask- and text-prompt, and provides higher-resolution images suitable for various image insertion tasks.

As shown in Table 1, compared to existing datasets (He et al. 2024; Choi et al. 2021), AnyInsertion covers diverse categories, offers higher resolution images, supports both mask- and text- prompts, and contains more samples.

Data Construction

Data Collection. Image insertion requires paired data: a reference image containing the element to be inserted, and a target image where the insertion occurs.

Concretely, we employ image matching techniques (Lindenberger, Sarlin, and Pollefeys 2023; Wang and Palpanas 2023) to create paired target and reference images and gather corresponding labels from web, leveraging the abundance of images showing accessories and people wearing them. For object data, we select images from MVImgNet (Yu et al. 2023), which provides varying viewpoints of common objects as reference-target pairs. For person insertion, we apply head pose estimation (Cobo et al. 2024) to select frames with similar head poses but varied body poses from the HumanVid dataset (Wang et al. 2024), which offers high-resolution video frames in real-world scenes. Frames with excessive motion blur are filtered out using blur detection (Pech-Pacheco et al. 2000), yielding high-quality data.

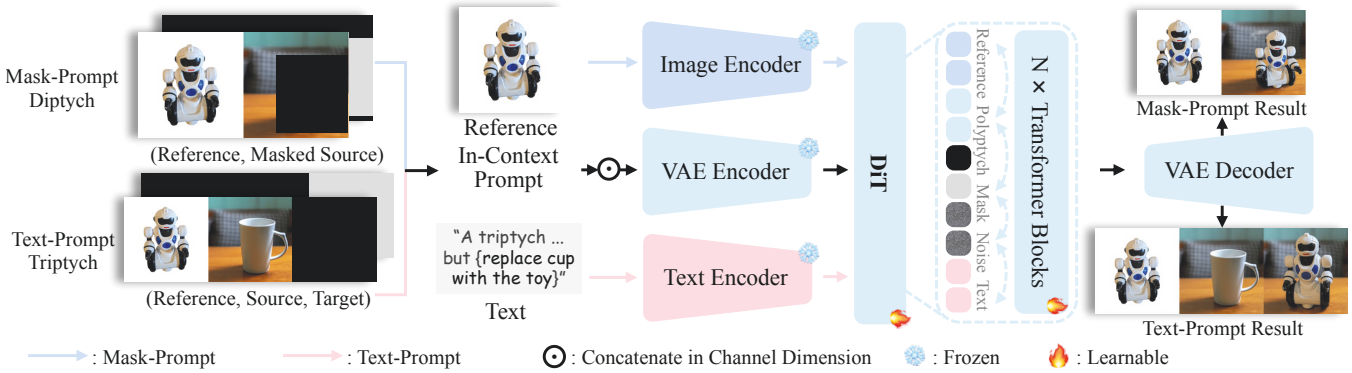


Figure 4: Overview of the Insert Anything model framework. Given different types of prompts, our unified framework processes in-context inputs (reference, source, and masks) through a frozen VAE encoder to preserve high-frequency details, and extracts semantic guidance from image and text encoders. These embeddings are combined and fed into learnable DiT transformer blocks for in-context learning, enabling precise and flexible image insertion guided by either mask- or text-prompt.

Data Generation. As shown in Fig. 3a, AnyInsertion provides data for both mask- and text-prompt control modes. Mask-prompt editing requires a mask to specify the insertion region in the target image, using elements from the reference image to fill the masked area of the target image. Text-prompt editing requires text to describe how the reference image's elements are inserted into the source image to form the target image. The data pairs for mask-prompt and text-prompt editing differ in structure.

Mask-Prompt. We use Grounded-DINO (Liu et al. 2024) and Segment Anything (SAM) (Kirillov et al. 2023) to generate reference and target masks from images and labels.

Text-Prompt. The source image, text description, and reference mask are derived from the reference and target images.

Source Image Generation. Source images are generated via two operations on target images: replacement and removal. For replacements, we use an editing LoRA fine-tuned from the FLUX.1 Fill [dev] model to produce initial edits. For removals, we use an object-removal LoRA model fine-tuned from FLUX.1 Fill [dev] to cleanly remove specified regions.

Text Creation. For replacement operations, we adapt instruction templates to reflect the desired transformations, such as *"replace `[source]` with `[reference]`"* for the transition from source to target. For addition operations, we use the format *"add `[label]`"* to describe the transition.

Reference Mask Extraction. We extract reference masks using the same method as in Mask-Prompt Editing.

More details on dataset construction can be found in §D in the supplementary material.

Dataset Overview

AnyInsertion dataset consists of training and testing subsets. The training set includes 136,385 samples across two prompt types: 58,188 mask-prompt image pairs (reference images, reference masks, target images, and target masks) and 78,197 text-prompt image pairs (reference images, reference masks, source images, target images, and texts). As shown in Fig. 3b, the dataset covers diverse categories including human subjects, daily necessities, garments, furniture, and various objects. This diversity enables the dataset to supports multiple insertion tasks including person insertion, object insertion, and garment insertion, thereby supporting a wide range of real-world applications. For evaluation, we curated a test set consisting of 158 data pairs: 120 mask-prompt pairs and 38 text-prompt pairs. The mask-prompt subset includes 40 pairs for object insertion, 30 pairs for garment insertion, and 60 pairs for person insertion (30 simple scene insertions and 30 complex scene insertions). The text-prompt subset contains 16 pairs for object insertion and 22 pairs for garment insertion.

Insert Anything Model

Overview. The image insertion task requires three key inputs: a reference image containing the element to be inserted, a source image providing the background context, and a control prompt (either mask or text) that guides the insertion process. The goal is to generate a target image that seamlessly integrates the element from the reference image (hereafter referred to as the "reference element") into the source image while preserving the identity of reference element (i.e., the visual features that define the reference element), and adhering to the specifications in the prompt. As illustrated in Fig. 4, our approach integrates three components: (1) a in-context format that organizes inputs to leverage contextual relationships, (2) semantic guidance mechanisms that extract high-level information from either text prompts or reference images, and (3) a DiT-based architecture that combines these elements through multimodal attention. Together, these components enable flexible control while maintaining visual harmony between inserted elements and their surrounding context.

In-Context Editing

In-context editing involves integrating reference elements into a source image while maintaining the contextual relationships between them. To achieve this, we first perform a background removal step to isolate the reference element. Following the approach (Chen et al. 2024b; Shin et al.

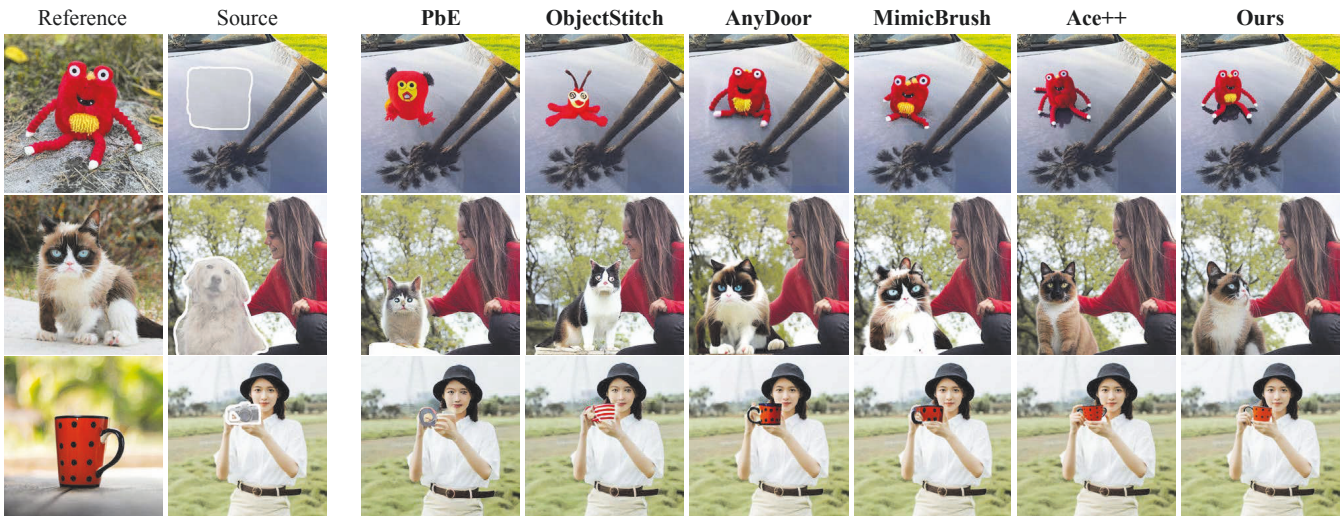


Figure 5: Qualitative comparison of mask-prompt object insertion results. Ours preserves object identity and visual coherence better than prior methods (Paint-by-Example (PbE), ObjectStitch, AnyDoor, MimicBrush, Ace++).

2024), we utilize the background removal process R_{seg} using Grounding-DINO and SAM to remove the background of the reference image, leaving only the object to be inserted.

Once the reference element is obtained, we perform in-context editing using two distinct approaches, corresponding to the mask-prompt and text-prompt modes.

`Mask-Prompt Diptych.` For mask-prompted editing, we propose a two-panel structure (diptych), which concatenates the processed reference image with a partially masked source image:

$$I_{\text{diptych}} = [R_{\text{seg}}(I_{\text{ref}}); I_{\text{masked\_src}}], \tag{1}$$

where I_{ref} represents the reference image and $I_{\text{masked\_src}}$ is the source image with its insertion region masked. We complement this visual input with a binary mask M_{diptych} that designates the reference image region (left panel) with zeros and the insertion region (right panel) with ones:

$$M_{\text{diptych}} = [\mathbf{0}_{h\times w}; M], \tag{2}$$

where $\mathbf{0}_{h\times w}$ has the same dimensions as each panel and M represents the insertion region. This structure provides clear spatial guidance while maintaining contextual relationships between the reference and target regions.

`Text-Prompt Triptych.` For text-prompted editing, we employ a three-panel structure (triptych) consisting of the processed reference image, the unmodified source image, and a fully masked region to be filled:

$$I_{\text{triptych}} = [R_{\text{seg}}(I_{\text{ref}}); I_{\text{src}}; \emptyset], \tag{3}$$

where I_{src} represents the source image and $\emptyset$ is the empty region to be generated. Similarly, we create a corresponding binary mask M_{triptych} that marks the reference and source regions with zeros and the generation region with ones:

$$M_{\text{triptych}} = [\mathbf{0}_{h\times w}; \mathbf{0}_{h\times w}; \mathbf{1}_{h\times w}], \tag{4}$$

where each component has the same dimensions as its corresponding panel.

Multiple Control Modes

Our framework supports two control modes for image insertion: mask-prompt and text-prompt. These modes enable flexible, task-specific editing by allowing users to specify insertion regions either manually through masks or via textual descriptions. To seamlessly integrate these input modalities, we leverage the multimodal attention mechanism of DiT (Peebles and Xie 2023), utilizing two dedicated branches: an image branch and a text branch.

In our framework, the image branch handles visual inputs, including the reference image, source image, and corresponding masks. These inputs are encoded into feature representations and concatenated with noise along the channel dimension to prepare for generation. In parallel, the text branch encodes the textual description to extract semantic guidance for image editing. The outputs from both branches are then fused via multimodal attention, enabling the model to jointly attend to visual and textual cues. This integration is formalized as:

$$Q = [Q_t; Q_i], K = [K_t; K_i], V = [V_t; V_i], \tag{5}$$

$$\text{MMA}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{6}$$

where $[;]$ represents the concatenation operation, and Q, K, and V are the query, key, and value components of the attention mechanism. The following describes how the attention mechanism operates under the two control modes.

Mask-Prompt. In mask-prompted editing, the insertion region in the source image is specified using a binary mask. This mask, along with the VAE-processed diptych, is concatenated with noise along the channel dimension and fed into the image branch of the DiT model. Simultaneously, semantic features from the reference image are extracted using a CLIP image encoder and passed into the text branch to provide contextual guidance.

Figure 6: Qualitative comparisons with AnyEdit and OminiGen on text-prompt object and garment insertion.

Figure 7: Mask-prompt garment insertion results.

Text-Prompt. In text-prompted editing, the insertion is guided by a textual description. The reference image informs the desired modifications, while the text prompt specifies the changes. The source image is modified accordingly to reflect the changes described in the text. To implement this, we design a specialized prompt template: *"A triptych with three side-by-side images. On the left is a photo of* `[label]`*; on the right, the scene is exactly the same as in the middle but* `[instruction]` *on the left."* This structured prompt provides semantic context, where the `[label]` identifies the reference element type, and the `[instruction]` specifies the modification. The input is processed through the text encoder, which provides guidance to the text branch of DiT. The triptych structure is processed by VAE and fed into the image branch, and the text token is concatenated with image features to enable joint attention between branches.

Experiments

Experimental Setup

Implementation Details. Our method builds upon FLUX.1 Fill [dev], a DiT-based inpainting model. It uses a T5 (Raffel et al. 2020) text encoder and a SigLIP (Zhai et al. 2023) image encoder, fine-tuned using LoRA (rank 256). For training, we use a batch size of 8 for mask prompts and 6 for text prompts, with all images at 768×768 resolution. We employed the Prodigy optimizer (Mishchenko and Defazio 2023) with safeguard warmup and bias correction enabled, applying a weight decay of 0.01. All experiments were conducted on 4 NVIDIA A800 GPUs (80GB each). Our AnyInsertion dataset served as our primary training set. We trained the model for 5000 steps across both prompt types (mask and text). For the sampling process, we performed denoising over 50 iterations. Our training loss function follows the flow matching (Lipman et al. 2022).

Test Datasets. We evaluate our method on three datasets: AnyInsertion, DreamBooth (Ruiz et al. 2023), and VTON-HD (Choi et al. 2021). From AnyInsertion, we select 40 samples for object insertion, 30 for garment insertion, and 30 for person insertion (simple scenes). For DreamBooth, we construct a test set of 30 group images, where one image serves as the reference and another as the target. We also evaluate on VTON-HD, the standard benchmark for virtual try-on and garment insertion.

Metrics. In experiments, evaluation metrics include Structural Similarity Index (Wang et al. 2004), Learned Perceptual Image Patch Similarity (Zhang et al. 2018), and Fréchet Inception Distance (Heusel et al. 2017).

Methods	AnyInsertion (Object)			DreamBooth		
	SSIM ↑	LPIPS ↓	FID ↓	SSIM ↑	LPIPS ↓	FID ↓
PbE	0.7612	0.0867	68.92	0.6110	0.1352	122.59
ObjectStitch	0.7386	0.1148	83.81	0.5709	0.1876	141.49
AnyDoor	0.7648	0.1831	67.99	0.5898	0.3029	95.14
MimicBrush	0.7371	0.2178	67.19	0.6039	0.2849	88.59
ACE++	0.6922	0.1485	40.11	0.5695	0.1823	64.39
Ours	**0.8791**	**0.0820**	**28.31**	**0.7820**	**0.1350**	**47.09**

Table 2: Quantitative comparison on *mask-prompt* object insertion tasks. On the AnyInsertion and DreamBooth datasets, Insert Anything outperforms existing methods.

Methods	AnyInsertion (Object)		
	SSIM ↑	LPIPS ↓	FID ↓
OminiGen	0.5558	0.3334	177.17
AnyEdit	0.5488	0.3473	226.25
Ours	**0.6678**	**0.2011**	**95.90**

Table 3: Quantitative comparison on *text-prompt* object insertion. On the AnyInsertion dataset, Insert Anything outperforms baselines (OminiGen, AnyEdit) across all metrics.

Methods	AnyInsertion (Garment)			VTON-HD		
	SSIM ↑	LPIPS ↓	FID ↓	SSIM ↑	LPIPS ↓	FID ↓
Ootdiffusion	0.8151	0.0970	87.38	0.8643	0.0605	28.36
CatVTON	0.8477	0.0607	36.62	0.8903	0.0513	24.80
Ours	**0.8665**	**0.0522**	**28.54**	**0.9161**	**0.0484**	**19.51**

Table 4: Quantitative comparison on mask-prompt garment insertion. Insert Anything achieves the best SSIM, LPIPS, and FID on both datasets.

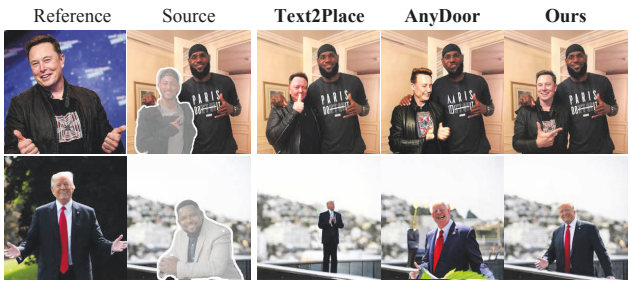


Figure 8: Mask-prompt person insertion results.

Figure 9: Ablation study on mask prompt insertion. Omitting in-context editing (IC Editing), semantic guidance, or AnyInsertion data degrades visual fidelity, sacrificing fine details and semantic cues.

Comparisons with Existing Works

We evaluate Insert Anything against state-of-the-art methods on object, garment, and person insertion tasks.

Object insertion. For mask-prompt insertion, we compare to Paint-by-Example (Yang et al. 2023), AnyDoor (Chen et al. 2024b), MimicBrush (Chen et al. 2024a), Ace++ (Mao et al. 2025), and our reproduction of ObjectStitch (Song et al. 2022). For text-prompt insertion, we include OmniGen (Xiao et al. 2025) and AnyEdit (Yu et al. 2024). As shown in Table 2, Insert Anything improves SSIM from 0.7648 to 0.8791 on AnyInsertion and from 0.6039 to 0.7820 on DreamBooth. In the text-prompt setting (Table 3), LPIPS decreases from 0.3473 to 0.2011, indicating better perceptual quality. Fig. 5 and 6 demonstrate that compared to prior methods, our approach handles challenging cases such as reflections on toys and car surfaces, interactions between animals and people, and composing a person with a cup, while preserving identity and achieving seamless integration.

Garment insertion. We benchmark against OOTDiffusion (Xu et al. 2024a) and CatVTON (Chong et al. 2024). Table 4 shows that on AnyInsertion we achieve the highest SSIM and lowest LPIPS and FID, and on VTON-HD we reduce LPIPS by 6%. Fig. 7 highlights our ability to preserve logos and text on garments and to adapt to large shape changes (e.g. skirt→trousers) with more realistic results.

Person insertion. We compare to Text2Place (Parihar et al. 2024) and AnyDoor (Chen et al. 2024b). Table 5 demonstrates that Insert Anything cuts LPIPS and FID by over 50% on AnyInsertion. Fig. 8 demonstrates that Insert Anything more effectively maintains facial identity and background consistency, yielding smoother interactions both between people and between a person and their surroundings.

Methods	AnyInsertion (Person)		
	SSIM ↑	LPIPS ↓	FID ↓
Text2Place*	0.8109	0.2641	138.26
AnyDoor	0.6807	0.3613	217.17
Ours	**0.8457**	**0.1269**	**52.77**

Table 5: Quantitative comparison on person insertion. * indicates the method requires both mask and text prompts.

Methods	SSIM ↑	LPIPS ↓	FID ↓
w/o In-Context Edting	0.8495	0.1095	59.97
w/o Semantic Guidance	0.8635	0.0876	39.63
w/o AnyInsertion	0.8483	0.1129	61.03
Ours	**0.8653**	**0.0865**	**35.72**

Table 6: Ablation results on mask-prompt insertion.

Ablation Study

We conducted an ablation study on mask-prompt image insertion. The values in Table 6 are weighted averages over person, object, and garment insertion, with weights matching their proportions in the mask-prompt test set (4:3:3 for object, garment, and person). Additional ablation results are provided in the supplementary.

In-Context Editing. As shown in Fig. 9, when we remove the polyptych in-context editing during training, the generated images fail to retain fine-grained details (e.g., textures) from the reference image and produces less natural person–cup interactions. This results in a significant drop in SSIM, and LPIPS values in Table 6, demonstrating the effectiveness of polyptych in-context editing in preserving high-frequency details and maintaining realistic compositing.

Semantic Guidance. As shown in Fig. 9, when we remove the semantic guidance for the reference image during training, the generated images lose high-level semantic information (e.g., color) from the reference image. This indicates that semantic guidance plays a crucial role in retaining the coarse, high-level semantic features.

AnyInsertion. When we remove our custom training data and rely solely on a training-free model for inference, Table 6 shows a noticeable drop in all evaluation metrics. Moreover, Fig. 9 illustrates that the model's ability to preserve facial details in person insertion is compromised.

Conclusion

This paper introduces Insert Anything, a unified framework for reference-based image insertion that overcomes the limitations of specialized approaches by supporting both mask- and text-guided control across diverse insertion tasks. Built on AnyInsertion, a large-scale dataset with 136K prompt–image pairs, we leverage DiT's multimodal modeling to implement an in-context editing mechanism with diptych and triptych prompts that preserves identity and visual coherence. Extensive experiments on three benchmarks show that our method outperforms state-of-the-art methods across person, object, and garment insertion, establishing a strong baseline for real-world applications.

Acknowledgments

This work was supported in part by the Fundamental Research Funds for the Central Universities (226-2025-00080) and the National Natural Science Foundation of China (U2336212). This work was also supported by the National Natural Science Foundation of China (62441617).

References

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18392–18402.

Chen, X.; Feng, Y.; Chen, M.; Wang, Y.; Zhang, S.; Liu, Y.; Shen, Y.; and Zhao, H. 2024a. Zero-shot Image Editing with Reference Imitation. *arXiv preprint arXiv:2406.07547*.

Chen, X.; Huang, L.; Liu, Y.; Shen, Y.; Zhao, D.; and Zhao, H. 2024b. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6593–6602.

Chen, X.; Zhang, Z.; Zhang, H.; Zhou, Y.; Kim, S. Y.; Liu, Q.; Li, Y.; Zhang, J.; Zhao, N.; Wang, Y.; et al. 2024c. UniReal: Universal Image Generation and Editing via Learning Real-world Dynamics. *arXiv preprint arXiv:2412.07774*.

Choi, S.; Park, S.; Lee, M.; and Choo, J. 2021. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14131–14140.

Chong, Z.; Dong, X.; Li, H.; Zhang, S.; Zhang, W.; Zhang, X.; Zhao, H.; and Liang, X. 2024. Catvton: Concatenation is all you need for virtual try-on with diffusion models. *arXiv preprint arXiv:2407.15886*.

Cobo, A.; Valle, R.; Buenaposada, J. M.; and Baumela, L. 2024. On the representation and methodology for wide and short range head pose estimation. *Pattern Recognition*, 149: 110263.

Han, Z.; Jiang, Z.; Pan, Y.; Zhang, J.; Mao, C.; Xie, C.; Liu, Y.; and Zhou, J. 2024. ACE: All-round Creator and Editor Following Instructions via Diffusion Transformer. *arXiv preprint arXiv:2410.00086*.

He, R.; Ma, K.; Huang, L.; Huang, S.; Gao, J.; Wei, X.; Dai, J.; Han, J.; and Liu, S. 2024. Freeedit: Mask-free reference-based image editing with multi-modal instruction. *arXiv preprint arXiv:2409.18071*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Kulal, S.; Brooks, T.; Aiken, A.; Wu, J.; Yang, J.; Lu, J.; Efros, A. A.; and Singh, K. K. 2023. Putting people in their place: Affordance-aware human insertion into scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17089–17099.

Li, X.; Yang, Z.; Quan, R.; and Yang, Y. 2024. Drip: Unleashing diffusion priors for joint foreground and alpha prediction in image matting. *Advances in Neural Information Processing Systems*, 37: 79868–79888.

Lindenberger, P.; Sarlin, P.-E.; and Pollefeys, M. 2023. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17627–17638.

Lipman, Y.; Chen, R. T.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

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.

Mao, C.; Zhang, J.; Pan, Y.; Jiang, Z.; Han, Z.; Liu, Y.; and Zhou, J. 2025. Ace++: Instruction-based image creation and editing via context-aware content filling. *arXiv preprint arXiv:2501.02487*.

Mishchenko, K.; and Defazio, A. 2023. Prodigy: An expeditiously adaptive parameter-free learner. *arXiv preprint arXiv:2306.06101*.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Ostrek, M.; O'Sullivan, C.; Black, M. J.; and Thies, J. 2024. Synthesizing Environment-Specific People in Photographs. In *European Conference on Computer Vision*, 292–309. Springer.

Parihar, R.; Gupta, H.; VS, S.; and Babu, R. V. 2024. Text2Place: Affordance-Aware Text Guided Human Placement. In *European Conference on Computer Vision*, 57–77. Springer.

Pech-Pacheco, J. L.; Cristóbal, G.; Chamorro-Martinez, J.; and Fernández-Valdivia, J. 2000. Diatom autofocusing in brightfield microscopy: a comparative study. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 3, 314–317. IEEE.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text

transformer. *Journal of machine learning research*, 21(140): 1–67.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

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.

Shen, K.; Quan, R.; Miao, J.; Xiao, J.; and Yang, Y. 2025. TarPro: Targeted Protection against Malicious Image Editing. *arXiv preprint arXiv:2503.13994*.

Shin, C.; Choi, J.; Kim, H.; and Yoon, S. 2024. Large-Scale Text-to-Image Model with Inpainting is a Zero-Shot Subject-Driven Image Generator. *arXiv preprint arXiv:2411.15466*.

Song, Y.; Zhang, Z.; Lin, Z.; Cohen, S.; Price, B.; Zhang, J.; Kim, S. Y.; and Aliaga, D. 2022. Objectstitch: Generative object compositing. *arXiv preprint arXiv:2212.00932*.

Song, Y.; Zhang, Z.; Lin, Z.; Cohen, S.; Price, B.; Zhang, J.; Kim, S. Y.; Zhang, H.; Xiong, W.; and Aliaga, D. 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.

Tan, Z.; Liu, S.; Yang, X.; Xue, Q.; and Wang, X. 2024. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 3.

Wang, Q.; and Palpanas, T. 2023. Seanet: A deep learning architecture for data series similarity search. *IEEE TKDE*, 35(12): 12972–12986.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wang, Z.; Li, Y.; Zeng, Y.; Fang, Y.; Guo, Y.; Liu, W.; Tan, J.; Chen, K.; Xue, T.; Dai, B.; et al. 2024. Humanvid: Demystifying training data for camera-controllable human image animation. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Xiao, S.; Wang, Y.; Zhou, J.; Yuan, H.; Xing, X.; Yan, R.; Li, C.; Wang, S.; Huang, T.; and Liu, Z. 2025. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13294–13304.

Xu, Y.; Gu, T.; Chen, W.; and Chen, C. 2024a. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. *arXiv preprint arXiv:2403.01779*.

Xu, Y.; Yang, Z.; and Yang, Y. 2025. SKDream: Controllable Multi-view and 3D Generation with Arbitrary Skeletons. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 314–325.

Xu, Y.; et al. 2024b. Gg-editor: Locally editing 3d avatars with multimodal large language model guidance. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 10910–10919.

Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; and Wen, F. 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.

Yang, Y.; Zhuang, Y.; and Pan, Y. 2021. Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies. *Frontiers of Information Technology & Electronic Engineering*, 22(12): 1551–1558.

Yu, Q.; Chow, W.; Yue, Z.; Pan, K.; Wu, Y.; Wan, X.; Li, J.; Tang, S.; Zhang, H.; and Zhuang, Y. 2024. AnyEdit: Mastering Unified High-Quality Image Editing for Any Idea. *arXiv preprint arXiv:2411.15738*.

Yu, X.; Xu, M.; Zhang, Y.; Liu, H.; Ye, C.; Wu, Y.; Yan, Z.; Zhu, C.; Xiong, Z.; Liang, T.; et al. 2023. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9150–9161.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11975–11986.

Zhang, K.; Mo, L.; Chen, W.; Sun, H.; and Su, Y. 2023. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36: 31428–31449.

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, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zhou, D.; Li, Y.; Ma, F.; Yang, Z.; and Yang, Y. 2024. Migc++: Advanced multi-instance generation controller for image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.