A^2-Edit: Precise Reference-Guided Image Editing of Arbitrary Objects and Ambiguous Masks

Huayu Zheng[1], Guangzhao Li[1,2], Baixuan Zhao[1], Siqi Luo[1], Hantao Jiang[1], Guangtao Zhai[1], and Xiaohong Liu[1,2⋆]

[1] Shanghai Jiao Tong University, Shanghai, China
[2] Shanghai Innovation Institute, Shanghai, China
{zhenghuayu, bxzhao7, Siqiluo647, eric_jht}@sjtu.edu.cn
{zhaiguangtao, xiaohongliu}@sjtu.edu.cn, gzhaoli03@gmail.com

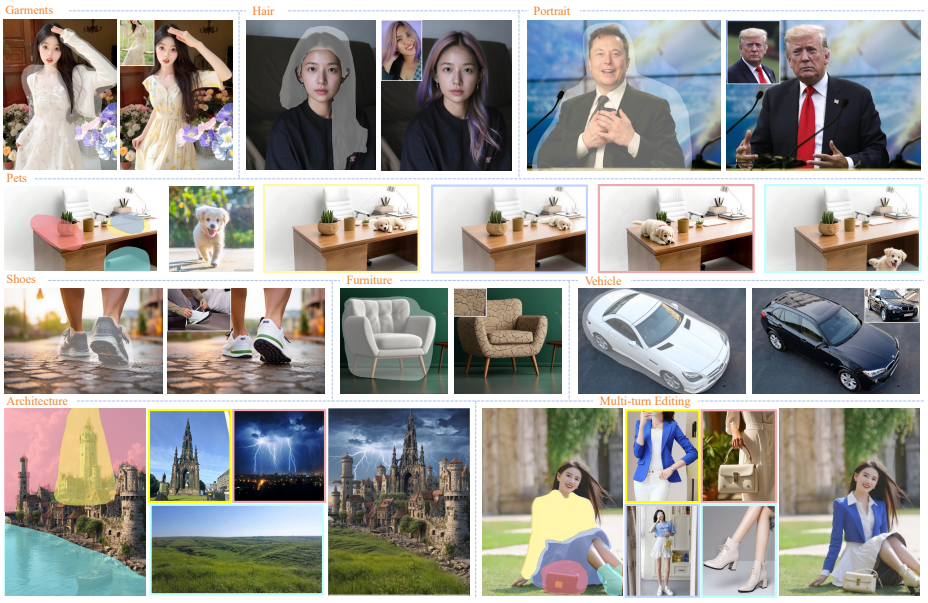

Fig. 1: Visual Results of A^2-Edit: A Unified Framework for Arbitrary-Object Inpainting. Our method supports a wide spectrum of real-world scenarios, delivering high-quality results across diverse object categories. The examples illustrate the generality and robustness of our method in handling various image editing tasks. More results are provided in the supplementary material.

Abstract. We propose **A^2-Edit**, a unified inpainting framework for arbitrary object categories, which allows users to replace any target region with a reference object using only a coarse mask. To address the issues of severe homogenization and limited category coverage in existing datasets, we construct a large-scale, multi-category dataset **UniEdit-500K**, which includes 8 major categories, 209 fine-grained subcategories, and a total of 500,104 image pairs. Such rich category diversity poses new challenges for

⋆ Corresponding author.

the model, requiring it to automatically learn semantic relationships and distinctions across categories. To this end, we introduce the **Mixture of Transformer** module, which performs differentiated modeling of various object categories through dynamic expert selection, and further enhances cross-category semantic transfer and generalization through collaboration among experts. In addition, we propose a **Mask Annealing Training Strategy** (MATS) that progressively relaxes mask precision during training, reducing the model's reliance on accurate masks and improving robustness across diverse editing tasks. Extensive experiments on benchmarks such as VITON-HD and AnyInsertion demonstrate that A^2-Edit consistently outperforms existing approaches across all metrics, providing a new and efficient solution for arbitrary object editing. The code is released at `https://github.com/huayu-zheng/A2Edit`.

Keywords: Reference-guided image editing · Image inpainting · Arbitrary object editing · Mask-guided generation

1 Introduction

In recent years, the rapid advancement of diffusion models [13,22,37,57] has significantly propelled the progress of image generation technologies. However, effectively adapting these models to precise and controllable image editing tasks [2, 26,27,39,44,48,53,54,56,61] remains challenging. Among various editing paradigms, reference-guided image inpainting has emerged as an important direction due to its strong controllability and practical value. This task requires transferring semantic content from a reference image to a masked target region while preserving object identity and maintaining natural consistency with the surrounding scene. Such capability forms the core technical foundation for applications including e-commerce product placement, virtual try-on, and personalized image customization.

Despite significant progress in recent years, existing reference-guided image inpainting methods still face critical bottlenecks in practical applications, which severely limit their generalization ability and usability. On the one hand, current approaches are typically optimized for specific object domains (e.g., garments [7, 55] or portraits [17, 36]), while editing tasks across different categories exhibit fundamentally different modeling objectives. For instance, garment editing emphasizes texture consistency, portrait editing focuses on identity preservation and morphological variations, and rigid objects require accurate geometric and material representation. Such disparities in objectives make it difficult for a unified model to satisfy cross-category optimization requirements through a single processing pathway. However, most existing methods still employ shared parameter pathways to process inputs from all categories, lacking category-specific modeling mechanisms, which leads to significant cross-category performance degradation or even failure. Meanwhile, existing datasets exhibit highly homogeneous category distributions and limited coverage [3, 4, 6, 12, 46], further weakening the model's ability to learn universal representations and exacerbating the generalization bottleneck.

On the other hand, current methods generally rely on high-precision segmentation masks to strictly define editing regions [3, 4, 46]. In real-world scenarios, however, whether using user-drawn sketches or automatically generated detection boxes, achieving pixel-level accuracy is difficult. When mask quality deteriorates, model performance degrades significantly. Such reliance on idealized spatial guidance not only reduces interaction flexibility but also makes models fragile in open environments, thereby hindering their practical deployment and application.

To address the aforementioned challenges, we propose **A^2-Edit**, a unified reference-guided image inpainting framework designed for **A**rbitrary object categories and **A**rbitrary levels of mask precision. A^2-Edit enhances both cross-category generalization and mask robustness within a unified model architecture through dynamic expert modeling and a progressive training mechanism. Specifically, we introduce a **Mixture of Transformers (MoT)** architecture, which dynamically routes input features to specialized expert subnetworks based on the semantic characteristics and category attributes of the input object. This design enables category-specific modeling while facilitating cross-category knowledge collaboration, thereby significantly improving the generalization capability of the unified model. Meanwhile, we propose the **Mask Annealing Training Strategy (MATS)**, which progressively reduces mask precision during training. This strategy encourages the model to shift from relying on strict geometric boundaries toward developing stronger contextual semantic understanding and structural generation capabilities, thereby improving its adaptability to coarse masks.

To support the effective training of this unified framework and mitigate the issue of homogeneous data distribution, we further construct a large-scale multi-category dataset, **UniEdit-500K**. The dataset contains over **500,000** reference–target image pairs, covering **8 major categories** and **209 fine-grained subcategories**, spanning diverse object types including texture-dominant non-rigid objects and structure-dominant rigid objects. Such diversified object distribution provides critical support for collaborative modeling under heterogeneous editing objectives and enables strong cross-category generalization within a unified model.

In summary, our main contributions are:

- We propose A^2-Edit, a highly generalized reference-guided image inpainting framework that breaks the constraints of domain-specific editing and rigid spatial guidance.
- We introduce a novel Mixture of Transformers (MoT) architecture and a Mask Annealing Training Strategy (MATS), enabling superior cross-category semantic transfer and robust performance with arbitrarily imprecise masks.
- We construct UniEdit-500K, the most comprehensive dataset to date for this task, covering 8 major domains and 209 subcategories to support universal editing research. Extensive evaluations show that our approach achieves state-of-the-art performance across multiple datasets and objective metrics.

2 Related Works

2.1 Image Editing

Diffusion models have surpassed traditional generative models (VAEs, GANs) [25, 49, 52] in image quality, establishing themselves as the mainstream choice for editing [8, 23, 31, 34, 43, 51, 59]. Early U-Net-based methods [1, 2, 13, 32, 55] often suffer from editing failures and detail distortions despite fine-grained masks. While Diffusion Transformers (DiTs) [10, 18, 37] have improved fidelity and controllability, recent text-driven approaches (IC-Edit [62], DreamO [33], X2I [29], Qwen-Image [50], FlowDirector [24]) still struggle with precise detail control. Mask-guided methods like Insert Anything [46] and ACE++ [30] achieve strong performance on rigid objects, yet remain inadequate for non-rigid subjects (e.g., human faces, pets) due to complex deformations and identity sensitivity. We propose A^2-Edit, a reference-based local editing framework that employs multi-precision mask fine-tuning to enable high-quality, fine-grained editing of diverse objects at arbitrary regions.

2.2 Universal Image Editing

Recent works have explored unified editing frameworks: ACE++ [30] unifies inputs via concatenated feature maps; Insert Anything expands training data across humans, objects, and garments; IC-Edit simulates in-context learning textually; UniReal [5] models editing as discontinuous video generation. However, these methods remain predominantly confined to rigid-object editing and struggle with non-rigid subjects requiring strong identity preservation. Text-driven approaches [29, 33] fail to maintain subject identity, while mask-guided methods [3, 4, 12, 30, 36] are constrained by mask precision. We propose A^2-Edit with an Mixture of Transformers (MoT) architecture that dynamically activates specialized experts based on semantic content and deformation characteristics, enabling a unified model that preserves structural stability for rigid objects and identity consistency for non-rigid subjects.

3 Method

3.1 Overview

The goal of image editing is to generate high-quality results in the target image by seamlessly integrating the subject from the reference image into the target scene, while adhering to the user-provided mask constraints and preserving the subject's identity even when the mask is imprecise or rough. As illustrated in Figure 2, our method integrates two key components: (1) a Mixture-of-Transformers (MoT) editing module that dynamically routes input features to different expert networks for differentiated modeling of diverse objects (§3.2); and (2) a Mask Annealing Training Strategy (MATS) that gradually reduces mask precision to shift the model from relying on strict spatial constraints to developing stronger

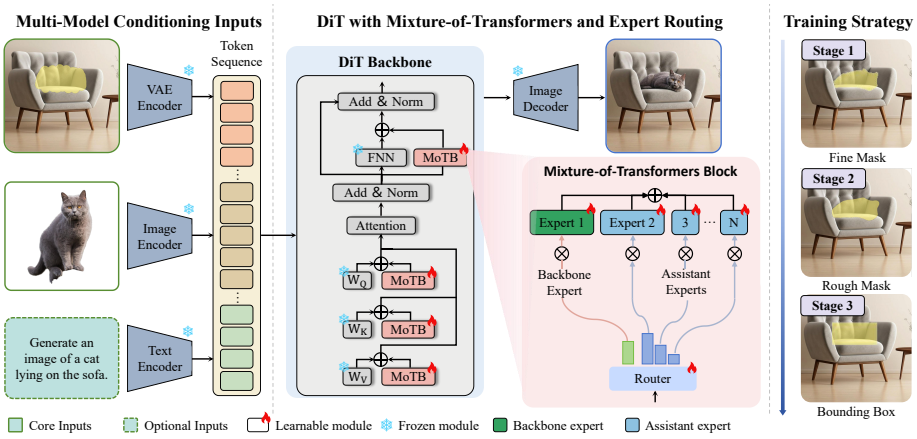

Fig. 2: Overview of the A^2-Edit framework. Our architecture takes a reference image, target image, and user mask as core inputs, encodes them into a unified feature space, and feeds the fused features into Mixture-of-Transformers (MoT) module. The MoT Block (MoTB) denotes an expert-specific Transformer block embedded within the attention and feed-forward layers. The model then dynamically routes features to specialized experts for differentiated modeling. Model is trained via Mask Annealing Training Strategy (MATS), a three-stage training process. Final outputs are decoded by a VAE decoder for high-fidelity results.

contextual understanding and content generation capabilities (§3.3). Working jointly, these components enable the model to handle complex, multi-category image editing tasks by maintaining fine-grained details and identity features of the reference element, while producing high-quality, naturally integrated results even under relaxed mask constraints.

3.2 Mixture of Transormers

Existing image editing methods typically adopt a unified parameter pathway to process inputs from different object categories. However, such a design struggles to accommodate diverse editing objectives and often exhibits poor generalization to unseen categories. In reference-guided image editing, different object categories vary significantly in both semantic content and structural patterns. For example, garment editing emphasizes texture consistency and contour alignment, portrait editing requires identity preservation and fine-grained non-rigid deformation modeling, while animals, furniture, vehicles, and buildings involve different geometric structures, scales, and spatial layouts. Therefore, a single shared parameter pathway can hardly model all these category-dependent editing requirements in a unified manner. In recent years, the *Mixture-of-Experts* (MoE) paradigm [9, 11, 15] has been widely adopted in large-scale models, but sparse experts are typically introduced only in the Feed-Forward Network (FFN) layers, while the Attention modules remain shared.

In image editing tasks, different object types require not only *category-specific content transformations* (handled by FFN), but also *specialized relational modeling* (handled by Attention), such as shape alignment, texture consistency, and spatial coherence. Therefore, applying expert modeling only to the FFN layers is often insufficient to capture these structural relationships. Meanwhile, recent studies [58] further show that Attention itself contains functionally decomposable subspaces and can benefit from conditional expert selection.

Based on these observations, we adopt a **Mixture-of-Transformers (MoT)** architecture that extends expert routing from the conventional FFN-only design to both the Attention and Feed-Forward Network (FFN) modules. Specifically, we introduce **Mixture-of-Transformers Blocks (MoTB)** in each Transformer layer, where multiple expert adapters are conditionally activated through a routing mechanism, enabling category-aware parameter specialization. Each expert is implemented as a lightweight **LoRA adapter** [14], enabling efficient parameter specialization with minimal computational overhead.

Multimodal Feature Encoding. Given a reference image I_r, a target image I_t, and a mask M, we first extract the foreground of the reference image to obtain a clean subject representation $\tilde{I}_r$ [4, 45]. The reference image is encoded to obtain visual features F_r, while the target image I_t with mask guidance is encoded into latent features F_t. If a text prompt P is provided, it is encoded into textual features F_p. All features are projected into a shared multimodal embedding space and fused into a unified token sequence:

$$\mathbf{x} = \text{Fuse}(F_t, F_r, F_p). \tag{1}$$

Fuse$(\cdot)$ denotes token-level multimodal fusion, where features from different modalities are projected into the same hidden dimension.

Expert Routing. We employ a multilayer perceptron as the gating network $G(\cdot)$ to predict routing weights over N experts. To balance model capacity and efficiency, we adopt an Anchor-Guided Routing (AGR) strategy $\mathcal{T}_{AGR}(\cdot)$. A backbone expert is always activated to provide universal knowledge, while assistant experts are activated only if their routing weights exceed the backbone's weight (or the highest-weighted expert otherwise). The activated expert set is defined as $\mathcal{S} = \mathcal{T}_{AGR}(\mathbf{w}(\mathbf{x}))$.

MoT Layer with LoRA Experts. For each linear transformation W in the Transformer layer (e.g., W_Q, W_K, W_V in attention and W_1, W_2 in FFN), we use a shared backbone weight with expert-specific LoRA increments:

$$W = W^{(0)} + \sum_{i \in \mathcal{S}} w_i(\mathbf{x}) \Delta W_i, \quad \Delta W_i = A_i B_i. \tag{2}$$

where $W^{(0)}$ denotes the shared backbone parameters and ΔW_i represents the LoRA update of the i-th expert. Notably, routing is performed only once per layer, and the same routing weights are shared by both the Attention and FFN modules to ensure consistent modeling behavior.

Training and Inference. All routing parameters and LoRA experts are optimized jointly in an end-to-end manner. During training, routing gradients automatically encourage experts to specialize in different category patterns while preserving shared knowledge. During inference, only a small subset of experts (the backbone and a few assistant experts) are activated, enabling efficient sparse computation. For out-of-distribution inputs, the gating network selects suitable expert combinations according to feature similarity, improving generalization and robustness.

3.3 Mask Annealing Training Strategy

Many models overly rely on idealized inputs, limiting user interaction freedom and degrading performance with poor mask quality. We propose the Mask Annealing Training Strategy (MATS), which gradually reduces input mask precision to guide the model from relying on fine spatial constraints to stronger contextual reasoning and content generation—enabling adaptive adjustments across objects/scenarios. It includes three training stages: *Fine Mask Training*, *Augmented Rough Mask Training*, and *Bounding Box Training*.

Fine Mask Training. In the initial training stage, we use high-precision segmentation masks strictly aligned with the target object to indicate the editing region. The model is provided with the target image I_t, its high-precision mask $M_t \in \{0,1\}^{H\times W}$, and reference image I_r.

Augmented Rough Mask Training. To reduce overreliance on mask boundaries and enhance tolerance for rough inputs, the second stage introduces mask augmentation. First, isotropic morphological dilation is applied to the fine mask M_t:

$$M_t^{dil} = M_t \oplus B_a, \tag{3}$$

where $\oplus$ denotes dilation and B_a is a circular structuring element of radius a, simulating user-drawn edge extensions. The external contour set $\mathcal{C} = \{\mathbf{p}_i = (x_i, y_i)\}_{i=1}^{N}$ is then extracted from M_t^{dil}. Contour points undergo spatially correlated random displacement via Perlin noise [38] to emulate human drawing errors:

$$\Delta\mathbf{p}_i = \alpha \cdot \begin{bmatrix} \mathcal{N}(x_i \cdot s, y_i \cdot s) \\ \mathcal{N}((x_i+\delta)\cdot s, (y_i+\delta)\cdot s) \end{bmatrix}, \tag{4}$$

where α controls displacement magnitude, s is noise scale, and δ decouples axes. Disturbed points are:

$$\hat{\mathbf{p}}_i = \Pi_\Omega(\mathbf{p}_i + \Delta\mathbf{p}_i), \tag{5}$$

with Π_Ω projecting coordinates to the image domain. The augmented rough mask $\bar{M}_t$ is derived from $\hat{\mathcal{C}} = \{\hat{\mathbf{p}}_i\}$.

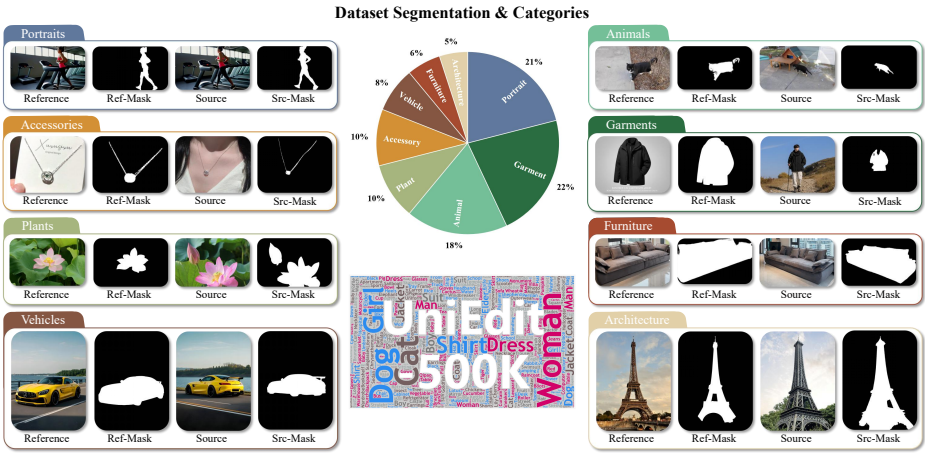

Fig. 3: An overview of our UniEdit-500K dataset. The central pie chart illustrates the proportional distribution across the eight major categories. The word cloud below visualizes the diversity of object classes within the dataset. Surrounding the center are representative examples from each category (Portraits, Accessories, Plants, Vehicles, Animals, Garments, Furniture, and Architecture), showcasing the data format which includes a reference image, a source image, and their corresponding segmentation masks.

Bounding Box Training. In the final stage, we further simplify the mask by extending the rough mask $\bar{M}_t$ to the target object's bounding box $\tilde{M}_t$. Without fine shape priors, the model infers the object's pose, scale, and local structure independently, generating reasonable, naturally integrated content with the reference image. This phase enhances creative generation capability and adaptability to objects (e.g., human faces, pets) with only rough localization.

4 UniEdit-500K

4.1 Comparison with Existing Datasets

Recent works have made remarkable progress in the field of image generation. However, most existing methods focus on single tasks, perform poorly on out-of-distribution object categories, struggling to adapt to complex and diverse real-world scenarios. We attribute this issue partly to the constrained categories in existing datasets: the FreeEdit [12] dataset mainly focuses on animals and plants, while the VITON-HD [6] dataset specializes in garments. Although AnyDoor [4] and MimicBrush [3] contain large scale data, they include very few samples related to human insertion. While AnyInsertion [46] expands categories to include humans, objects, and garments, it still falls short in category coverage. To address these limitations, we construct UniEdit-500K, a large-scale dataset

covering multiple domains and diverse object types. This dataset comprises eight major categories—garments, portraits, animals, plants, accessories, furniture, vehicles, and architecture—further refined into 209 fine-grained subcategories. The diversity of data categories enhances the model's comprehensive capabilities across various object types. More importantly, such diversity provides the model with abundant examples to learn both distinctions and correlations among different categories, thereby establishing essential conditions for achieving genuine generalization.

4.2 Data Construction

We employed a multi-strategy hybrid approach to jointly construct the large-scale UniEdit-500K dataset.

For **garments and accessories** categories, we build upon the VITON-HD and DressCode datasets, and additionally collect a large number of *person-wearing-item* scene images from open-licensed platforms such as Pixabay that allow research use and redistribution. Furthermore, we supplement the dataset with images sourced from community platforms under explicit creator permission. During data construction, we pair the standalone item images with their corresponding wearing-scene images to form high-quality image pairs.

For **portraits and animals**, which requires strong identity consistency, we adopted a video-driven dynamic sampling method. The video data is sourced from the VidGen-1M dataset [47] (Apache License 2.0). Given a video and a specified subject category, we first performed frame sampling and used Grounding DINO [28] and Segment Anything Model (SAM) [16] to extract subject masks; then, we computed quality scores via no-reference image quality assessment to filter high-quality frames; finally, we selected the optimal two frames based on the degree of subject variation as the output.

For **plants, vehicles, architecture, and furniture** objects, we collected multi-view images from and open data platforms, combining manual verification and automatic pairing strategies to construct image pairs with sufficient viewpoint variation.

In addition, we utilized NanoBanana to address missing data and balance the dataset for certain domains. After obtaining matched image pairs, we performed object detection using Grounding DINO with the obtained object category c, producing bounding boxes b_i. Then, based on b_i, we generated binary segmentation masks $M_i \in \{0,1\}^{H\times W}$ using the SAM. Ultimately, we obtained paired reference and target images along with their corresponding masks, as illustrated in Figure 3. For images obtained with creator authorization, we strictly restrict usage to the explicitly permitted subset to ensure compliance and proper data governance.

4.3 Dataset Overview

The UniEdit-500K dataset consists of a training subset and a test subset. The training set contains 500,104 samples covering eight major categories (clothing, portraits, animals, plants, accessories, furniture, vehicles, and buildings), comprising a total of 209 fine-grained subcategories. Each data sample includes a reference image, a target image, and the corresponding mask. The composition of the dataset is illustrated in Figure 3, with additional details provided in the supplementary materials. For evaluation, we constructed a test set of 200 image pairs, selecting 25 samples from each major category to comprehensively assess the model's capability across different object classes.

5 Experiments

5.1 Experimental Setup and Configuration

Implementation Details We adopt FLUX.1 Fill [19] as the backbone of our approach, which is an open-source image inpainting model based on the DiT architecture. The encoder utilizes a T5 text encoder [42] and a SigLIP image encoder [60]. The model is configured with LoRA of rank 32, 8 expert subnetworks, and the Anchor-Guided Routing strategy—where one backbone expert is always activated, and assistant experts are activated based on routing weights. During training, the batch size is set to 8, and all images are processed at a resolution of 1024×1024. All experiments are conducted on a cluster of 4 NVIDIA A100 GPUs. We use our self-constructed UniEdit-500K dataset as the main training set and train the model for 6000 steps in total, including 3000 steps with fine masks, 1500 steps with augmented rough masks, and 1500 steps with bounding box masks. In terms of computational cost, during inference on a single NVIDIA A100 (80GB), our model reaches a peak memory usage of 42 GB (42,074 MiB), which is slightly higher than the 40 GB (39,884 MiB) required by the baseline model. For 50 sampling steps, the average inference time of our model is approximately 30 seconds, compared to about 32 seconds for the baseline model (both results are averaged over 100 evaluation samples).

Evaluation Settings We evaluate our method on two standard datasets: VITON-HD [6] and AnyInsertion [46]. VITON-HD serves as the standard benchmark for virtual try-on and garment insertion tasks. The AnyInsertion test set contains 100 samples, including 40 object samples, 30 garment samples, and 30 portrait samples. In addition, to comprehensively validate cross-category generalization capability, we construct a test set of 200 samples by selecting 25 samples from each of the eight major categories in UniEdit-500K. For evaluation metrics, we adopt five quantitative measures: CLIP-I [41], DINO-I [35], LPIPS, FID, and a VLM-based score. CLIP-I, DINO-I, and LPIPS are used to assess semantic alignment and perceptual similarity from multiple perspectives, while FID evaluates the overall distributional realism of the generated images. Moreover, since metrics such as CLIP, DINO, and LPIPS do not always adequately capture boundary artifacts, visual realism, and local consistency, we further introduce a

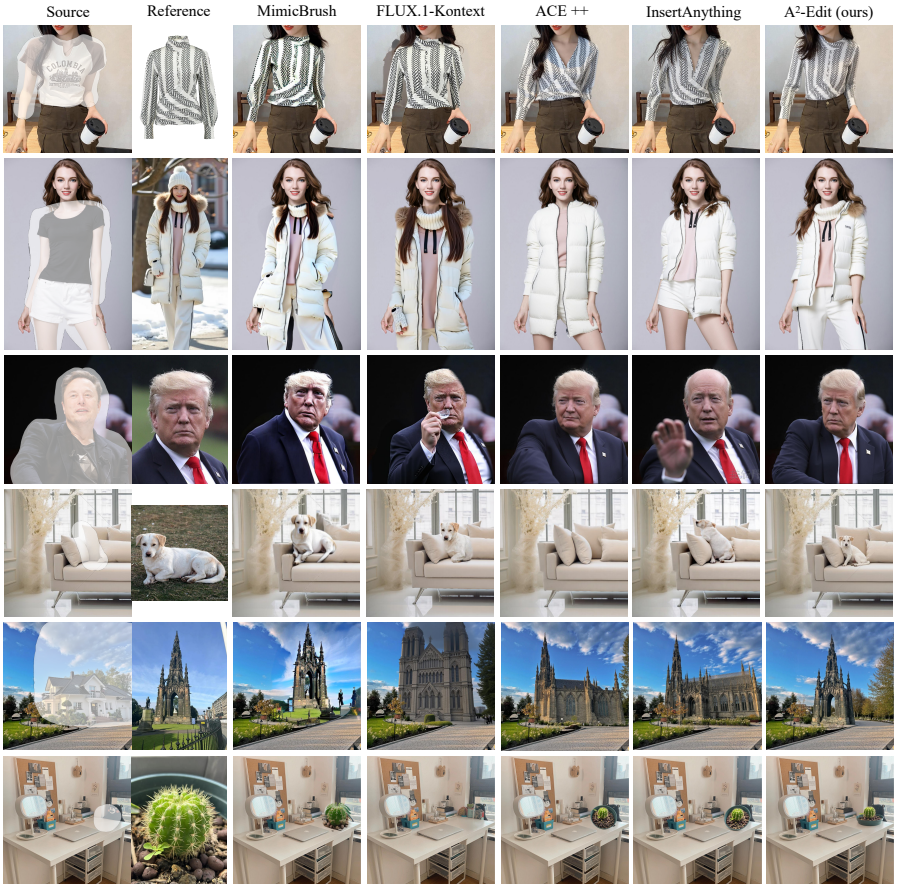

Fig. 4: Qualitative comparison with existing mask-guided image editing methods. Our method consistently produces more coherent structures, sharper details, and better semantic alignment with the target edit compared to existing methods(MimicBrush [3], FLUX.1-Kontext [20], ACE++ [30] and InsertAnything [46]).

Vision-Language Model (VLM)-based evaluation protocol. Specifically, we employ a large language model to score the generated results along three aspects: boundary quality, realism, and local consistency. The final VLM score is obtained by averaging these three dimensions across all samples. Detailed evaluation protocols and scoring prompts are provided in the supplementary material.

5.2 Quantitative Comparison

Common Domain Evaluation: In this study, we compare A^2-Edit with FLUX.1-Fill-dev [19], AnyDoor [4], MimicBrush [3], FLUX.1-Kontext-dev [20], ACE++ [30] and InsertAnything [46]) on the VITON-HD and AnyInsertion datasets, using both fine and rough masks. The rough masks are obtained by

Table 1: Quantitative comparison with existing image editing methods. Evaluations conducted on VTON-HD and AnyInsertion datasets show that our A^2-Edit model significantly outperforms existing methods. The best and second-best results are demonstrated in **bold** and underlined, respectively. Note that higher DINO-I and CLIP-I scores indicate better performance, while lower LPIPS values are preferred.

Method	VTON-HD						AnyInsertion					
	Fine Mask			*Rough Mask*			*Fine Mask*			*Rough Mask*		
	DINO-I	CLIP-I	LPIPS	DINO-I	CLIP-I	LPIPS	DINO-I	CLIP-I	LPIPS	DINO-I	CLIP-I	LPIPS
FLUX.1-Fill	58.59	72.59	0.1241	46.18	69.04	0.1509	50.37	66.06	0.1962	43.57	62.89	0.2673
AnyDoor	59.18	72.54	0.1270	45.99	69.38	0.1587	48.15	65.01	0.2070	42.47	60.67	0.2854
MimicBrush	58.99	73.01	0.1022	49.13	68.54	0.1479	50.29	68.88	0.1878	47.89	65.13	0.2011
FLUX.1-Kontext	60.09	79.44	0.0800	56.11	77.91	0.0972	59.49	74.70	0.1094	54.15	74.05	0.1384
ACE++	62.33	78.00	0.0767	58.78	78.33	0.0813	57.69	74.33	0.1241	56.44	73.80	0.1327
Insert Anything	62.15	79.96	0.0728	60.87	78.22	0.0809	60.90	76.11	0.0978	60.50	75.98	0.0992
A^2-Edit (Ours)	**64.07**	**81.53**	**0.0677**	**63.79**	**80.19**	**0.0685**	**61.67**	**77.45**	**0.0909**	**61.73**	**77.27**	**0.0903**

applying isotropic morphological dilation to the fine masks. As shown in Table 1, A^2-Edit achieves the best performance across all metrics, indicating that A^2-Edit has the ability to perform as well as specialized models in common data domain. Additional qualitative comparisons with representative text-guided editing models, including Qwen-Image-Edit-2509 [40] and FLUX.2-klein-4B [21], are provided in the supplementary material for cross-paradigm comparison.

Generalization Ability Evaluation: To test the model's ability in different object categories, we evaluate A^2-Edit against other methods on the UniEdit test set. As presented in Table 2, A^2-Edit significantly outperforms existing approaches on this multi-category dataset while maintaining comparable performance with those in common domains. This indicates that A^2-Edit exhibits more stable performance and stronger generalization capability when dealing with diverse object categories in real-world scenarios, showcasing superior universality.

User Study: In Figure 5, we showcase the outcomes of the results of our user study. A total of 24 participants were recruited to compare the images generated by our method with those produced by various baseline approaches. The reported percentages indicate the proportion of participants who preferred the results of a given model in each comparison. The result demonstrates that our method consistently receives higher preference from participants compared to all evaluation methods.

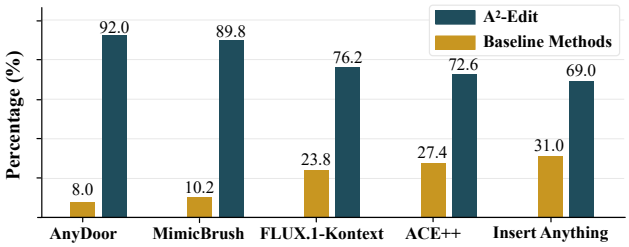

Fig. 5: Statistical results of user study.

Table 2: Quantitative comparison on the UniEdit test set across diverse object categories. Higher VLM scores indicate better performance, while lower FID values are preferred.

Method	UniEdit									
	Fine Mask					*Rough Mask*				
	DINO-I ↑	CLIP-I ↑	LPIPS ↓	FID ↓	VLM ↑	DINO-I ↑	CLIP-I ↑	LPIPS ↓	FID ↓	VLM ↑
FLUX.1-Fill	45.61	63.69	0.2519	93.70	42.7	40.09	61.58	0.3739	128.91	26.3
AnyDoor	46.76	64.23	0.2534	92.19	41.0	40.42	61.39	0.3623	120.80	28.3
MimicBrush	50.02	68.56	0.2133	86.23	47.7	45.07	64.36	0.2871	100.54	36.3
FLUX.1-Kontext	53.47	73.99	0.1403	70.54	52.3	48.29	70.56	0.2091	80.94	54.0
ACE++	<u>58.13</u>	74.70	0.1164	62.30	76.0	<u>56.68</u>	<u>73.66</u>	<u>0.1219</u>	<u>63.19</u>	74.7
Insert Anything	56.22	<u>74.87</u>	<u>0.1121</u>	<u>59.99</u>	<u>78.7</u>	56.13	73.37	0.1243	61.48	<u>75.0</u>
A^2-Edit (Ours)	**61.71**	**76.89**	**0.0898**	**57.72**	**82.3**	**62.28**	**77.45**	**0.0897**	**54.62**	**79.3**

5.3 Qualitative Comparison

Figure 4 presents a comprehensive qualitative comparison between A^2-Edit and MimicBrush, FLUX.1-Kontext, ACE++, and InsertAnything across a variety of inpainting-based image editing tasks, including garments, portraits, pets, and other object categories. All experiments are conducted using hand-drawn rough masks. In several challenging editing scenarios, MimicBrush shows clear limitations in preserving fine details and fails to maintain identity consistency in portrait editing. FLUX.1-Kontext exhibits instability when handling mask boundaries, often producing noticeable edge artifacts and visible seams. InsertAnything and ACE++ frequently struggle to correctly interpret editing intentions under coarse mask conditions, generating results that remain closer to the original image style while suffering from missing details or indiscriminate filling within the masked region.

In contrast, A^2-Edit is able to accurately capture editing objectives across diverse object categories and complex scenes, producing semantically consistent, structurally complete, and visually stable results. For example, in garment editing tasks, it preserves the structural integrity of clothing while generating consistent texture details; in portrait replacement tasks, it maintains strong identity consistency; and in architectural replacement tasks, it effectively handles the blending between foreground and background regions. These results demonstrate the strong cross-category generalization ability of our unified editing framework. More qualitative results are provided in the supplementary material.

5.4 Ablation Study

A^2-Edit framework When we gradually remove the A^2-Edit framework and use standard MoE in NLP that only route FFNs or the baseline model for training, the test results are shown in Figure 6. The overall generation quality of the model decreases across all tested categories, leading to a significant drop in all quantitative metrics in Table 3. The standard MoE shows a noticeable improvement over the baseline model, but it still remains significantly lower than our method. This observation further validates the critical role of the A^2-Edit framework in achieving cross-category generalization.

Table 3: Ablation studies. We evaluate model variants on UniEdit. w/o denotes *without*, and w. denotes *with*. AE: Arbitrary Editing framework; MoE: MoE applied only to FFN layers; UE: UniEdit-500K; FT: fine-mask training; RT: augmented rough-mask training; BT: bounding-box-mask training.

Method	*Fine Mask*					*Rough Mask*				
	DINO-I ↑	CLIP-I ↑	LPIPS ↓	FID ↓	VLM ↑	DINO-I ↑	CLIP-I ↑	LPIPS ↓	FID ↓	VLM ↑
w/o. AE	56.32	72.45	0.1423	75.37	64.3	56.18	73.21	0.1364	72.63	65.7
w. MoE	56.69	74.01	0.1211	71.90	70.3	56.11	73.69	0.1366	73.04	70.3
w/o. UE	58.64	74.32	0.1182	70.04	68.7	56.27	73.01	0.1347	72.84	66.3
w. FT	60.27	75.83	0.0981	67.55	71.0	55.65	72.92	0.1516	74.77	64.7
w. FT + RT	61.44	76.43	0.0941	63.72	74.7	59.70	74.86	0.1073	60.33	73.0
Top-2 router	57.78	74.96	0.1168	62.49	72.7	57.42	75.90	0.1101	58.87	72.7
Fixed-2 router	56.30	73.89	0.1335	68.63	68.3	55.21	73.49	0.1406	73.13	66.0
softmax router	60.00	75.61	0.1034	58.41	77.0	60.95	76.91	0.0983	56.12	74.7
seperate router	60.56	75.72	0.0979	58.56	78.0	61.16	77.34	0.1003	54.90	77.3
Ours	**61.71**	**76.89**	**0.0898**	**57.72**	**82.3**	**62.28**	**77.45**	**0.0897**	**54.62**	**79.3**

Fig. 6: A^2-Edit Ablation Study. Removing the framework, training data, or MATS reduces generation quality and cross-category generalization, while progressively adding each component significantly improves detail fidelity and the model's ability to handle rough masks.

Anchor-Guided Routing To further validate the routing design in A^2-Edit, we compare AGR with several alternative routing strategies, including Top-K routing, Fixed-K routing, and softmax routing. Top-K routing activates the K experts with the highest routing scores, Fixed-K routing always activates the same K experts, and softmax routing uses normalized expert weights without hard selection. We set $K = 2$ to match the number of activated experts in AGR. We also compare our shared routing design with separate routers for Attention and FFN, where the two modules independently select experts. As shown in Table 3, AGR achieves the best performance, confirming the benefit of combining a stable backbone expert with an adaptive assistant expert. The backbone expert preserves general editing priors across categories, while the assistant expert dynamically captures category-specific structural and semantic variations. Overall, these ablations validate the critical role of the A^2-Edit framework, including joint Attention-FFN expert routing and Anchor-Guided Routing, in achieving robust cross-category generalization.

UniEdit-500K As shown in Table 3, when our constructed UniEdit-500K training data is removed and only the untrained initial model is used for inference, all evaluation metrics exhibit a significant decline. Although the model retains a certain level of cross-category generation capability through the framework itself, it demonstrates notable deficiencies in detail generation quality and task intent comprehension. These results highlight the essential value of the UniEdit-500K dataset in enhancing the model's generalization ability and achieving cross-category object editing.

Mask Annealing Training Strategy To validate the effectiveness of MATS, we compare the model performance across four training stages: without training stage (without UniEdit-500K), training with only fine masks, training with fine masks and augmented rough masks, and the full A^2-Edit. As shown in Figure 6, when the model is trained only with fine masks, it tends to generate content excessively strictly aligned with the mask boundaries during inference with fine masks, and fails to perform meaningful edits when given rough masks due to its inability to interpret the intended editing region. After incorporating rough-mask training, this issue is significantly alleviated. Furthermore, as reported in Table 3, adding bbox mask training leads to additional improvements, indicating that MATS plays a crucial role in reducing the model's dependence on precise masks and enhancing its stability and generalization across diverse object categories. More ablation results are provided in the supplementary material.

6 Conclusion

This paper presents A^2-Edit, a unified reference-guided image inpainting framework that overcomes the limitations of specialized approaches by supporting editing tasks across arbitrary object categories and mask precision levels. Building upon our newly constructed UniEdit-500K dataset, we implement the dynamically routed Mixture of Transformers architecture and Mask Annealing Training Strategy, effectively preserving identity consistency while enhancing the model's adaptability to diverse editing requirements and real-world input conditions. Extensive experiments across multiple benchmarks demonstrate that **A^2**-Edit consistently achieves state-of-the-art performance across various categories, establishing a new technical standard for reference-guided image editing and providing a versatile solution for practical creative applications.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62301310, 62572317 and 62225112.

References

1. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
2. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)
3. Chen, X., Feng, Y., Chen, M., Wang, Y., Zhang, S., Liu, Y., Shen, Y., Zhao, H.: Zero-shot image editing with reference imitation. Advances in Neural Information Processing Systems **37**, 84010–84032 (2024)
4. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6593–6602 (2024)
5. Chen, X., Zhang, Z., Zhang, H., Zhou, Y., Kim, S.Y., Liu, Q., Li, Y., Zhang, J., Zhao, N., Wang, Y., et al.: Unireal: Universal image generation and editing via learning real-world dynamics. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12501–12511 (2025)
6. Choi, S., Park, S., Lee, M., Choo, J.: Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14131–14140 (2021)
7. Chong, Z., Dong, X., Li, H., Zhang, S., Zhang, W., Zhang, X., Zhao, H., Liang, X.: Catvton: Concatenation is all you need for virtual try-on with diffusion models. arXiv preprint arXiv:2407.15886 (2024)
8. Cui, X., Wu, G., Gan, Z., Zhai, G., Liu, X.: Face2qr: A unified framework for aesthetic, face-preserving, and scannable qr code generation. Advances in Neural Information Processing Systems **37**, 29142–29165 (2024)
9. Dai, D., Deng, C., Zhao, C., Xu, R., Gao, H., Chen, D., Li, J., Zeng, W., Yu, X., Wu, Y., et al.: Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. arXiv preprint arXiv:2401.06066 (2024)
10. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
11. Han, X., Wei, L., Dou, Z., Sun, Y., Han, Z., Tian, Q.: Vimoe: An empirical study of designing vision mixture-of-experts. IEEE Transactions on Image Processing **34**, 7209–7221 (2025)
12. He, R., Ma, K., Huang, L., Huang, S., Gao, J., Wei, X., Dai, J., Han, J., Liu, S.: Freeedit: Mask-free reference-based image editing with multi-modal instruction. arXiv preprint arXiv:2409.18071 (2024)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems (2020)
14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
15. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural computation **3**(1), 79–87 (1991)
16. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)

17. Kulal, S., Brooks, T., Aiken, A., Wu, J., Yang, J., Lu, J., Efros, A.A., Singh, K.K.: Putting people in their place: Affordance-aware human insertion into scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17089–17099 (2023)
18. Labs, B.F.: Flux. https://github.com/black-forest-labs/flux (2024)
19. Labs, B.F.: Flux.1-fill-dev. https://huggingface.co/black-forest-labs/FLUX.1-Fill-dev (2024)
20. Labs, B.F.: Flux. https://huggingface.co/black-forest-labs/FLUX.1-Kontext-dev (2025)
21. Labs, B.F.: Flux. https://huggingface.co/black-forest-labs/FLUX.2-klein-4B (2025)
22. Li, C., Wu, H., Hao, H., Zhang, Z., Kou, T., Chen, C., Liu, X., BAI, L., Lin, W., Zhai, G.: G-refine: A general refiner for text-to-image generation. In: ACM Multimedia 2024 (2024)
23. Li, G., Cen, K., Zhao, B., Xin, Y., Luo, S., Zhai, G., Zhang, L., Liu, X.: LayerT2V: A unified multi-layer video generation framework. arXiv preprint arXiv:2508.04228 (2025)
24. Li, G., Yang, Y., Song, C., Liu, X., Zhang, C.: Flowdirector: Training-free flow steering for precise text-to-video editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7805–7815 (2026)
25. Li, W., Wu, G., Wang, W., Ren, P., Liu, X.: Fastllve: Real-time low-light video enhancement with intensity-aware look-up table. In: ACM MM (2023)
26. Li, X., Yang, Z., Quan, R., Yang, Y.: Drip: Unleashing diffusion priors for joint foreground and alpha prediction in image matting. Advances in Neural Information Processing Systems **37**, 79868–79888 (2024)
27. Liu, D., Xin, Y., Zhao, S., Zhuo, L., Lin, W., Li, X., Qin, Q., Zhai, G., Liu, X., Li, H., et al.: Lumina-mgpt: Flexible photorealistic autoregressive text-to-image generation. International Journal of Computer Vision **134**(4), 141 (2026)
28. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European Conference on Computer Vision. pp. 38–55. Springer (2024)
29. Ma, J., Peng, Q., Guo, X., Chen, C., Lu, H., Yang, Z.: X2i: Seamless integration of multimodal understanding into diffusion transformer via attention distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16733–16744 (2025)
30. Mao, C., Zhang, J., Pan, Y., Jiang, Z., Han, Z., Liu, Y., Zhou, J.: Ace++: Instruction-based image creation and editing via context-aware content filling. arXiv preprint arXiv:2501.02487 (2025)
31. Maximov, M., Elezi, I., Leal-Taixé, L.: Ciagan: Conditional identity anonymization generative adversarial networks. In: CVPR (2020)
32. Meng, C., Ho, J., Chen, Y., Song, J., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. In: ICLR (2021)
33. Mou, C., Wu, Y., Wu, W., Guo, Z., Zhang, P., Cheng, Y., Luo, Y., Ding, F., Zhang, S., Li, X., et al.: Dreamo: A unified framework for image customization. arXiv preprint arXiv:2504.16915 (2025)
34. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)

35. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
36. Parihar, R., Gupta, H., VS, S., Babu, R.V.: Text2place: Affordance-aware text guided human placement. In: European Conference on Computer Vision. pp. 57–77. Springer (2024)
37. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
38. Perlin, K.: An image synthesizer. In: Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH). pp. 287–296 (1985)
39. Qin, Q., Zhuo, L., Xin, Y., Du, R., Li, Z., Fu, B., Lu, Y., Li, X., Liu, D., Zhu, X., et al.: Lumina-image 2.0: A unified and efficient image generative framework. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20031–20042 (2025)
40. Qwen: qwen. `https://github.com/QwenLM/Qwen-Image` (2025)
41. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763 (2021)
42. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research **21**(140), 1–67 (2020)
43. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
44. Shen, K., Quan, R., Zhu, L., Xiao, J., Yang, Y.: Audioscenic: Audio-driven video scene editing. arXiv preprint arXiv:2404.16581 (2024)
45. Shin, C., Choi, J., Kim, H., Yoon, S.: Large-scale text-to-image model with inpainting is a zero-shot subject-driven image generator. arXiv preprint arXiv:2411.15466 (2024)
46. Song, W., Jiang, H., Yang, Z., Quan, R., Yang, Y.: Insert anything: Image insertion via in-context editing in dit. arXiv preprint arXiv:2504.15009 (2025)
47. Tan, Z., Yang, X., Qin, L., Li, H.: Vidgen-1m: A large-scale dataset for text-to-video generation. arXiv preprint arXiv:2408.02629 (2024)
48. Wang, J., Pu, J., Qi, Z., Guo, J., Ma, Y., Huang, N., Chen, Y., Li, X., Shan, Y.: Taming rectified flow for inversion and editing. arXiv preprint arXiv:2411.04746 (2024)
49. Wang, W., Wu, G., Cai, W., Zeng, L., Chen, J.: Robust prior-based single image super resolution under multiple gaussian degradations. IEEE Access **8**, 74195–74204 (2020)
50. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
51. Wu, G., Liu, X., Jia, J., Cui, X., Zhai, G.: Text2qr: Harmonizing aesthetic customization and scanning robustness for text-guided qr code generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8456–8465 (2024)
52. Wu, G., Zhao, L., Wang, W., Zeng, L., Chen, J.: Pred: A parallel network for handling multiple degradations via single model in single image super-resolution. In: ICIP (2019)

53. Wu, G., Zheng, H., Luo, S., Zhai, G., Liu, X.: Animateqr: Bridging aesthetics and functionality in dynamic qr code generation. Advances in Neural Information Processing Systems **38**, 3472–3490 (2026)
54. Xin, Y., Qin, Q., Luo, S., Zhu, K., Yan, J., Tai, Y., Lei, J., Cao, Y., Wang, K., Wang, Y., et al.: Lumina-dimoo: An omni diffusion large language model for multimodal generation and understanding. arXiv preprint arXiv:2510.06308 (2025)
55. Xu, Y., Gu, T., Chen, W., Chen, C.: Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. arXiv preprint arXiv:2403.01779 (2024)
56. Xu, Y., Zhu, L., Yang, Y.: Gg-editor: Locally editing 3d avatars with multimodal large language model guidance. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 10910–10919 (2024)
57. Yang, Y., Liu, Z., Jia, J., Gao, Z., Li, Y., Sun, W., Liu, X., Zhai, G.: Diffstega: towards universal training-free coverless image steganography with diffusion models. arXiv preprint arXiv:2407.10459 (2024)
58. Yang, Y., Wang, C., Li, J.: Umoe: Unifying attention and ffn with shared experts. arXiv preprint arXiv:2505.07260 (2025)
59. Zhai, L., Guo, Q., Xie, X., Ma, L., Wang, Y.E., Liu, Y.: A3gan: Attribute-aware anonymization networks for face de-identification. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 5303–5313 (2022)
60. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
61. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. Advances in Neural Information Processing Systems **36**, 31428–31449 (2023)
62. Zhang, Z., Xie, J., Lu, Y., Yang, Z., Yang, Y.: In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. arXiv preprint arXiv:2504.20690 (2025)