SpatialDiff: 3D-Aware Object Movement via Implicit Spatial Modeling

Zheng Liu[1*] Zijian He[1] Huiguo He[2] Weizhi Zhong[1]
Yejun Tang[3] Huan Yang[3] Kun Gai[3] Guanbin Li[1,4,5†]
[1]Sun Yat-sen University [2]South China University of Technology [3]Kuaishou Technology
[4]Shenzhen Loop Area Institute [5]Guangdong Key Laboratory of Big Data Analysis and Processing
liuzh385@mail2.sysu.edu.cn, liguanbin@mail.sysu.edu.cn

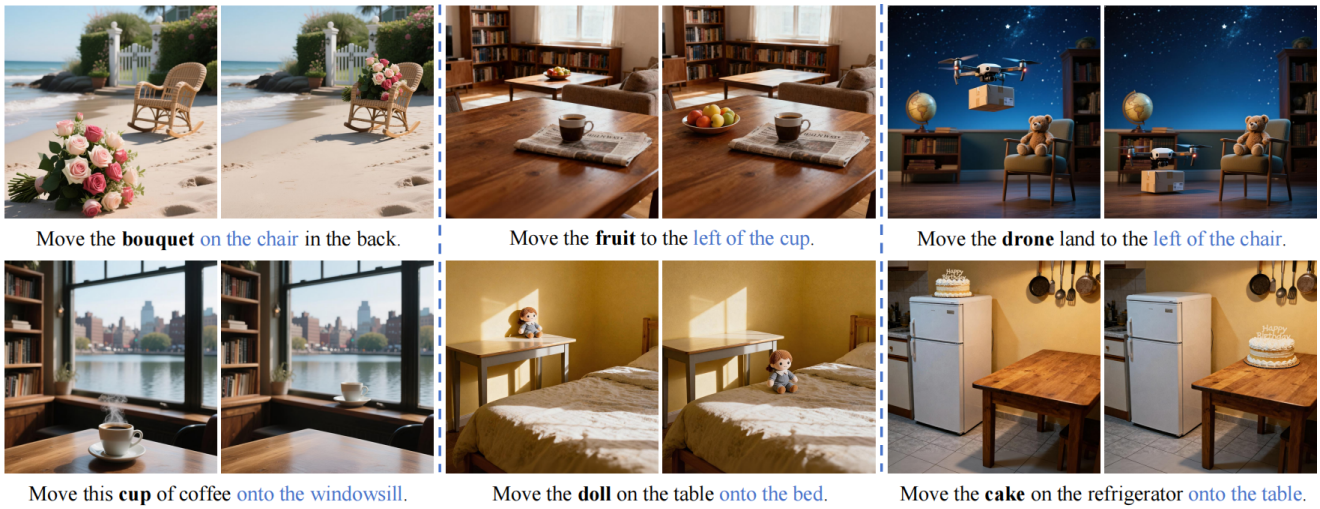


Figure 1. Results of our **Implicit Spatial Modeling** method. Our method captures complex relative positions (e.g., middle, front, top) and performs high-fidelity multi-dimensional transformations while maintaining semantic consistency, making it highly suitable for advanced scene content creation.

Abstract

Recent advances in image editing allow impressive manipulation of objects, existing methods still struggle to handle spatial movement in complex scenes, such as objects span different depth layers or are partially occluded. Most image editing methods focus solely on prior information from 2D datasets, emphasizing planar features while lacking support for spatial structures. Even approaches that incorporate explicit positional information fail to capture true 3D spatial relationships, thus limiting accurate object movement in complex scenes. In this paper, we present ***SpatialDiff****, a method that effectively captures 3D spatial structures, enabling precise and consistent object movements in complex scenes. Our core innovations are twofold: (1)* ***Implicit 3D Spatial Modeling****, which introduces 3D prior knowledge and enables the model to internally build a comprehensive understanding of the three-dimensional spatial structure; and (2)* ***Global Spatial Supervision****, which constrains the latent spatial features to enable the model to perceive changes in object spatial positions caused by editing operations. Experimental results demonstrate that our method significantly improves the accuracy and fidelity of spatial movement in complex scenes.*

1. Introduction

Recent years have seen rapid progress in the field of image editing [6, 7, 9, 35, 38, 46, 48–50], especially driven by advances in diffusion-based generative models [8, 17, 28, 31]. As the quality of base models improves, tasks such as style transfer [29, 38], object removal and insertion [1, 6, 11, 19], and fine-granular content editing [4, 5, 12, 48, 50] have become increasingly viable. In many real-world editing sce-

*Work done while interning at Kuaishou Technology.
†Corresponding author is Guanbin Li.

narios, users may modify an object's appearance, and relocating it within the scene based on a given instruction is often equally important. However, achieving precise and natural control over object movement within a single image remains challenging, as illustrated in Figure 2.

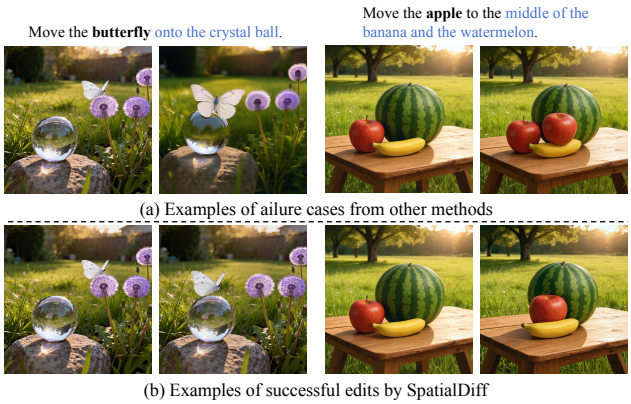
Methods [2, 13, 21, 32, 39, 40, 45, 47] that operate purely in the 2D domain, particularly those leveraging pre-trained diffusion models, provide a more straightforward and widely applicable solution. In these approaches, the editing process remains in image or latent space, the user issues an instruction (e.g., via text), and the model manipulates the source image accordingly. These instruction-driven methods benefit from ease of use and efficiency. Nonetheless, they share a fundamental drawback: in the absence of 3D spatial knowledge (depth, spatial layout), they struggle to guarantee that object motions are consistent in 3D space, especially under user instructions that demand precise spatial relocations.

A more natural solution to address this is via 3D-aware methods [3, 22, 26, 37, 44, 51]. By invoking explicit three-dimensional modelling or reconstruction, such methods aim to reason about object geometry, camera viewpoint, depth, occlusion and spatial relationships. For instance, Diffusion Handles [26] use the estimated depth map to lift diffusion activations for an object to 3D. Diff3DEdit [37] predict novel views of the selected object using estimated depth maps, and these novel views act as a geometry critic to correct misalignments in 3D shapes while editing. LACONIC [22] leverages the input 3D layout and camera pose to model explicit scene geometry and guide the repositioning of objects. In principle this grants strong geometric control and spatial consistency. However, when faced with the complex scene editing scenario, these approaches face substantial limitations. Explicit 3D reconstruction from a single image is inherently ill-posed: unknown viewpoint, depth ambiguities, occlusion, incomplete geometry and lack of multi-view supervision all degrade the reconstruction quality and thus impair the subsequent editing. As a result, though 3D-based pipelines hold promise, their applicability in 2D image editing remains limited.

Summarising the above: on the one hand, 2D-based diffusion editing methods are flexible and effective but lack spatial understanding and thus cannot always execute instruction-level spatial movement accurately; on the other hand, 3D-based editing methods promise spatial reasoning but falter on single-image inputs and complex scene conditions. This observation raises a key question:

(Q) Can we bring the benefits of 3D spatial priors into the 2D image-editing diffusion framework implicitly, rather than through full 3D reconstruction, so that the diffusion transformer (DiT) can understand and control object spatial positioning relations?

To address this, we propose **SpatialDiff**, an instruction-driven image editing framework that achieves flexible and precise spatial movement of objects in complex scenes as shown in Figure 1. SpatialDiff introduces an ***Implicit 3D Spatial Modeling*** mechanism by leveraging a 3D visual geometry encoder to extract geometry-aware features from the input image, and a Connector module to align these features with the DiT latent space. This design enables the model to internalize spatial priors and reason about 3D structure within a purely 2D diffusion process. Furthermore, to guide the model toward a coherent spatial understanding, we introduce a ***Global Spatial Supervision*** strategy. During training, the depth map of the target image is used as an auxiliary supervision signal to regularize the utilization of 3D spatial tokens. Through this design, our method bridges the gap between 2D and 3D paradigms, mitigating the challenges of single-image 3D reasoning while enriching diffusion-based editing with implicit spatial awareness, enabling precise and spatial consistent object movement. In summary, our key contributions are as follows:



(a) Examples of ailure cases from other methods

(b) Examples of successful edits by SpatialDiff

Figure 2. **(a)** Examples of failure cases from other methods: (1) fail to maintain natural appearance and 3D consistency; (2) fail to perform object movement correctly in complex scenes. **(b)** Examples of successful edits by SpatialDiff.

- We propose SpatialDiff, the first instruction-driven method that implicitly incorporates 3D prior information to enable precise spatial object movement.
- We introduce Global Spatial Supervision, a latent depth supervised training mechanism, which leverages the depth map of the target image during training to smoothly guide the model's spatial understanding.
- Experimental results demonstrate that our model achieves state-of-the-art performance in both instructional movement consistency and image quality preservation.

2. Related Work

2.1. Instruction-based Image Editing

Instruction-based image editing methods [10, 12, 15, 18, 21, 39, 41, 47, 50] enable flexible image modification by tak-

ing natural language commands as input, allowing users to achieve targeted edits on a reference image with improved controllability and a simplified workflow. FlowEdit [15] explicitly constructs a transformation path between the reference and target images; ChronoEdit [41] models the image editing process as a continuous video sequence. Methods such as SmartEdit [12], MGIE [10], and FireEdit [50] introduce vision-language models (VLMs) to assist in training diffusion models, thereby enhancing their understanding and reasoning capabilities for complex instructions. Furthermore, Flux-Kontext [18] significantly improves instruction alignment by scaling up both model parameters and training data. Step1X-Editing [21], BAGEL [47], and Qwen-Image-Edit [39] leverage multimodal large language models (MLLMs) to jointly model input images and user instructions, extracting latent embeddings that are concatenated with Gaussian noise, and then using the model's in-context learning capability to generate the target image. By incorporating this multimodal understanding, these methods enhance the model's editing capabilities in complex scenes while maintaining semantic consistency and visual fidelity. However, despite these advances, they remain limited by the lack of injected 3D knowledge, which prevents spatially coherent object motion under fine-grained movement instructions.

2.2. 3D-Aware Image Editing

Recent works have sought to address spatial inconsistency in image editing by incorporating explicit 3D reasoning into diffusion frameworks [16, 22, 24, 26, 27, 30, 37, 51]. These approaches introduce geometric awareness through reconstruction or spatial lifting, enabling models to reason about object structure, depth, and viewpoint. Diffusion Handles [26] lifts diffusion activations into 3D space using estimated depth maps for object manipulation. Diff3DEdit [37] synthesizes novel views guided by depth estimation to refine 3D shape alignment during editing. LACONIC [22] leverages scene layout and camera pose to model explicit geometry and guide object repositioning. In parallel, single-image 3D reconstruction methods [20, 34, 51] attempt to infer object shapes from sparse or single-view inputs, allowing the object to be temporarily "lifted" to 3D, manipulated, and then reprojected back to 2D. However, these approaches struggle to handle multiple objects and complex spatial relationships within a single image, where occlusion and layout ambiguity make consistent reconstruction difficult. Although 3D-aware designs can improve spatial consistency and geometric control, they remain constrained by the inherent challenges of single-image 3D inference. Their reliance on explicit reconstruction, while providing spatial reasoning capabilities, also limits their robustness and scalability in general 2D image editing scenarios.

3. Preliminary

Diffusion or flow-based models learn data distributions by progressively denoising Gaussian-corrupted samples. Their strong generative ability and controllable sampling make them widely used in instruction-driven image editing. InstructPix2Pix [4] is a pioneering work in instruction-driven image editing, which fine-tunes a pretrained diffusion model on a large-scale dataset of (input image, instruction, target image) triplets to modify specific objects or regions in an image x according to a user-provided natural language instruction I.

In this work, we adopt a flow-based formulation of diffusion, where the forward process linearly interpolates between a clean image and noise to generate intermediate samples:

$$x_t = (1-t)x + t\epsilon, \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, I)$ and $t \in [0, 1]$ denotes the noise level. For a flow-based model parameterized by θ, the training objective is defined as:

$$\mathcal{L} = \mathbb{E}_{x,t,c,\epsilon\sim\mathcal{N}(0,I)} \, \|v - v_\theta(x_t, t, c)\|_2^2, \tag{2}$$

where the target velocity is $v = \epsilon - x$. The denoising network v_θ can be conditioned on auxiliary information c, such as a text prompt, a reference image, or other control signals, enabling instruction-guided image editing.

4. Method

We introduce SpatialDiff, an implicit modeling approach that incorporates 3D spatial information from the input image into the image editing process. We detail Implicit 3D Spatial Modeling in Section 4.1, Global Spatial Supervision in Section 4.2, and the two-stage training strategy in Section 4.3.

4.1. Implicit 3D Spatial Modeling

Existing image editing methods typically handle object movement based on 2D prior knowledge, making them inadequate for complex scenes involve occlusion or depth variation. We argue that this limitation primarily stems from the model's lack of spatial understanding of the 3D world, and thus propose to incorporate Implicit 3D Spatial Modeling (ISM) into its internal representations.

3D Visual Geometry Encoder. To endow the diffusion backbone with a spatially grounded understanding, we extract spatial-aware features from the input image using a 3D Visual Geometry Encoder (3D-VGE). This encoder is based on the Transformer backbone of a recent 3D foundational model, VGGT [36], which directly infers all key 3D attributes of a scene, including camera parameters, point maps, depth maps, and 3D point tracks, from one image.

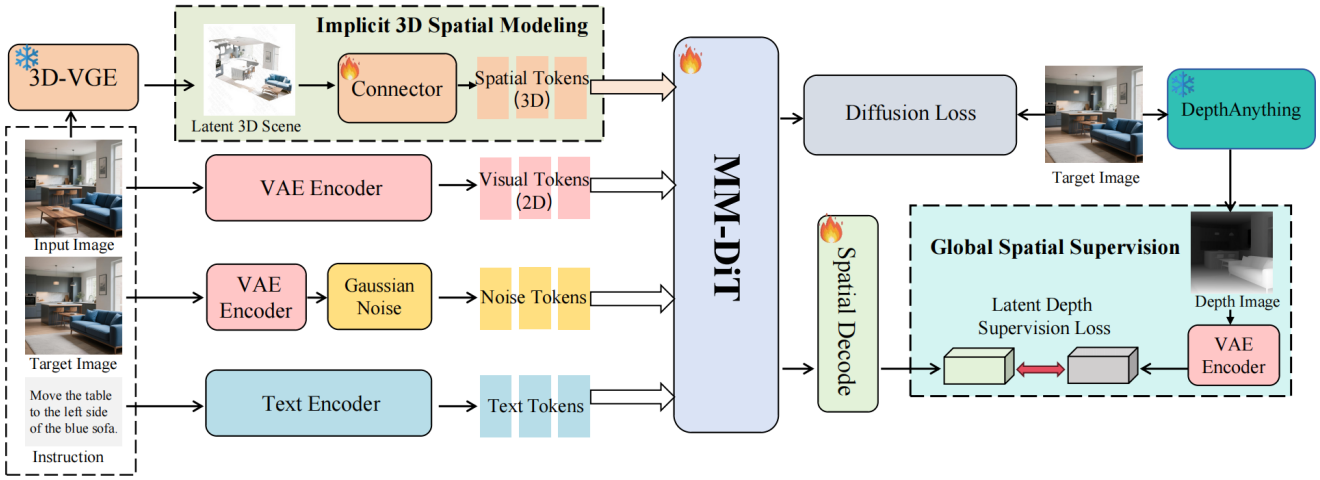


Figure 3. The pipeline of SpatialDiff. SpatialDiff leverages a single image along with its latent 3D scene representation to achieve precise control over object positions in 2D image scenes, while preserving the global spatial structure and semantic consistency.

Given an input image x_0, we obtain its spatial-aware representation as:

$$S = \text{3D-VGE}(x_0), \quad S \in \mathbb{R}^{l' \times d'}, \tag{3}$$

where l' denotes the token sequence length and d' the feature dimension of the spatial tokens. By leveraging the backbone features without using the task-specific prediction heads, 3D-VGE provides implicit 3D geometric priors to guide the image editing process, enabling the model to benefit from 3D spatial knowledge without explicitly reconstructing depth, surface normals, or other 3D attributes. However, since these extracted features differ from the DiT latent representation, an alignment module is required before they can be effectively integrated.

Connector for Spatial Alignment. We introduce a Connector module that aligns the 3D Visual Geometry Encoder features with the latent feature space of the diffusion transformer. The connector employs a set of learnable query tokens $Q_l \in \mathbb{R}^{l \times d}$, where l corresponds to the token sequence length in the DiT space and d denotes the token dimension. These query tokens selectively extract and align spatial information from the spatial features S through a cross-attention mechanism to match the latent representation of the DiT:

$$\hat{Q}_l = \text{Attn}(Q_l, S) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{4}$$

where

$$Q = Q_l W_Q^T, \quad K = SW_K^T, \quad V = SW_V^T. \tag{5}$$

Here, W_Q, W_K, W_V are learnable projection matrices of dimensions $\mathbb{R}^{d \times d}$, $\mathbb{R}^{d \times d'}$, and $\mathbb{R}^{d \times d'}$, respectively. Through this cross-attention alignment, the Connector module distills spatial-aware information into a representation that aligns well with the DiT latent space.

Token Integration. The aligned 3D spatial tokens are then concatenated with other conditional tokens to form the complete latent sequence:

$$C = [x, \hat{Q}_l, I], \tag{6}$$

where x denotes the latent tokens of the input image, Q_l denotes the aligned spatial tokens, and I the instruction tokens. This unified token sequence is then processed by the multimodal layers of the DiT backbone, where context learning integrates local appearance features, spatial cues, and instructions across the entire sequence. Through this multimodal interaction, SpatialDiff implicitly captures 3D spatial structures, including object depth, relative positioning, and inter-object spatial relationships, enabling precise and geometrically consistent manipulations even in complex scenes with occlusion or varying depth layers.

4.2. Global Spatial Supervision

Although implicit 3D spatial modeling introduces valuable 3D knowledge, we observe that even when an object is successfully relocated, traces may remain at its original position, or the overall semantic consistency of the scene may be affected (see Figure 5, Model B). This suggests that while the aligned 3D spatial features help the model comprehend object geometry and its initial position, they primarily reinforce static spatial memory rather than driving dynamic spatial updates. As a result, the model lacks a mechanism to dynamically guide the scene to remain consistent with the target image during object movement.

To address this issue, we introduce Global Spatial Supervision (GSS), which applies depth-guided supervision

during training to align the original 3D spatial tokens with the spatial positions of the target image. We explored two strategies for depth supervision: (1) Explicit Depth Supervision, which enforces pixel level consistency between decoded 3D spatial features and target depth maps in image space, and (2) Latent Depth Supervision, which constrains their correspondence in the VAE latent space. While the first strategy, an explicit approach, provides hard alignment, it often leads to unnatural editing results and causes the loss of content in non-edited regions (see Figure 5, Model C). Supervision in latent space emphasizes high level spatial and semantic relationships over low level pixel details, delivers a smooth and robust training signal less sensitive to subtle variations in detail, and enables effective integration of 3D spatial priors. Therefore, we adopt Latent Depth Supervision in this work, which promotes consistency in the latent geometry and better preserves relative spatial relationships among scene elements.

Formally, we obtain the target image's depth map using a pretrained monocular depth estimator [43]:

$$d_{\text{tgt}} = \text{DepthAnything}(x_{\text{tgt}}), \tag{7}$$

where x_{tgt} denotes the ground-truth edited image in the training pair.

Specifically, during training, we obtain $\hat{s}$, the DiT-processed spatial tokens that encode spatial aware information and map them into the VAE latent space via a learnable spatial decode head $\bar{\mathcal{D}}$. We then enforce consistency with the encoded target depth representation using a mean squared error (MSE) loss:

$$\mathcal{L}_{\text{GSS}} = \| \bar{\mathcal{D}}(\hat{s}) - \mathcal{E}(d_{\text{tgt}}) \|_2^2, \tag{8}$$

where $\mathcal{E}(\cdot)$ denotes the encoder of the VAE.

Unlike prior 3D-aware approaches that rely on explicit reconstruction or depth estimation during inference, our method leverages explicit depth information solely as a smooth auxiliary supervision signal in training. This design preserves the implicit nature of our spatial modeling while enabling the initial 3D information to be dynamically updated during training, encoding geometry features in a globally consistent manner. As a result, SpatialDiff achieves a more robust and coherent understanding of spatial structure, effectively suppressing residual artifacts while preserving implicit 3D reasoning without requiring any explicit 3D reconstruction at inference.

4.3. Training Strategy

To ensure stable convergence and fully exploit the capability of the Connector, we employ a two-stage training strategy. In the first stage, only the Connector module is optimized to align 3D spatial features with the DiT latent space, guided by the standard diffusion loss:

$$\mathcal{L}_{Align} = \mathbb{E}_{x_t,t,C,\epsilon\sim\mathcal{N}(0,I)} \|v - v_\theta(x_t, t, C)\|_2^2, \tag{9}$$

In the second stage, we jointly optimize the DiT backbone, the Connector module, and the spatial decoder head that projects the 3D spatial features into the VAE latent space. During this stage, the training objective incorporates Global Spatial Supervision to ensure that the model maintains consistency throughout the generation process:

$$\mathcal{L} = \mathbb{E}_{x_t,t,C,\epsilon\sim\mathcal{N}(0,I)} \|v - v_\theta(x_t, t, C)\|_2^2 + \lambda \cdot \mathcal{L}_{\text{GSS}}, \tag{10}$$

where λ = 0.01 is a weighting factor balancing the flow matching loss and the GSS loss.

5. Experiments

5.1. Experimental Setups

Implementation Details. We use Flux-Kontext [2] as the base model for instruction-driven image editing, and employ VGGT [36] as the 3D foundation model to extract 3D features from the reference image. To align the 3D spatial features with the DiT space, we construct the Connector module using 8 cross-attention layers, with the learnable query tokens sequence length set to 1024. The model training is based on the code provided by [42], with the LoRA rank set to 64. The training images have a resolution of 512 $\times$ 512. In all experiments, we set the learning rate to 1e-4 on 8 GPUs, employed AdamW with a batch size of 4 per GPU, and performed 3k steps of training.

Datasets and Evaluation Benchmarks. Due to the lack of datasets for instruction-driven spatial object movement in 2D image editing tasks, we adapted the OBJect-3DIT [23] movement dataset. Specifically, we collected the source and target images from OBJect-3DIT and used Qwen3-VL-32B-Instruct [33] to generate editing instructions describing the transformation from source to target images, thereby constructing a movement dataset containing 20k (reference image, instruction, target image) triplets. This dataset was used for training, without relying on any 3D asset information(e.g., coordinates). To more comprehensively evaluate the model's performance across different scenarios, we construct an additional complex scene benchmark **SpatialBench** containing 100 images with foreground, midground, and background objects. The comprehensive evaluation benchmark consists of 50 original OBJect-3DIT test images and 100 additionally constructed images depicting complex real-world scenes. We evaluate the two benchmarks separately, results on SpatialBench are presented in Sections 5.2 and 5.3, while results on the OBJect-3DIT test images are provided in the Supplementary Material.

Comparison Methods. We compared SpatialDiff with state-of-the-art instruction-driven image editing models, including Flux-Kontext [18], Step1X-Edit [21], OmniGen2 [40], BAGEL [47], and Qwen-Image-Editing [39]. Most of these methods employ a multimodal large language model (MLLM) to jointly encode the reference image and

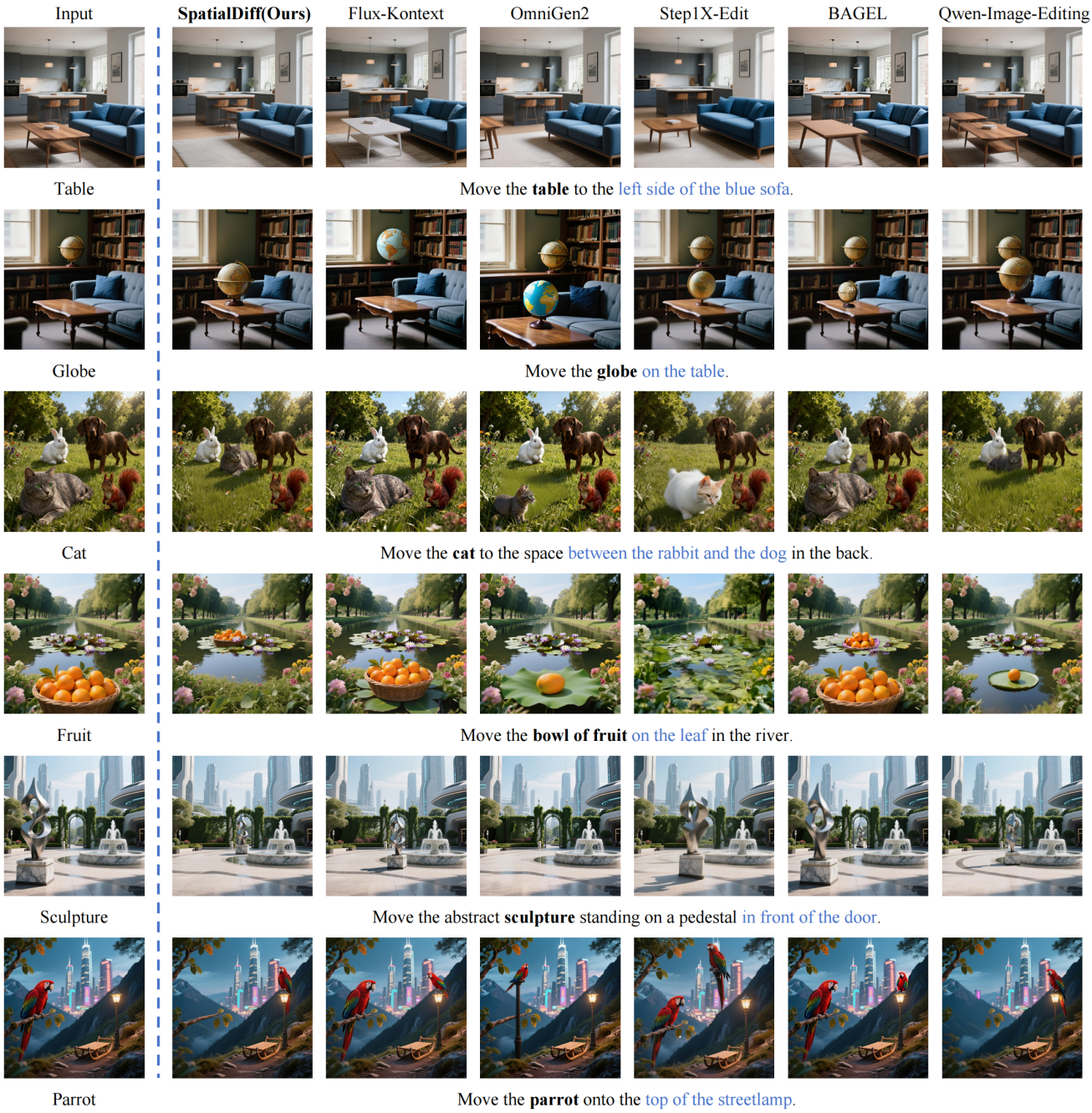


Figure 4. **Qualitative comparison**. The leftmost image represents the input image, with the target object to be edited shown below it. Each example is accompanied by the corresponding editing instruction, where the blue text indicates the spatial position involved in the editing operation.

the editing instruction, enabling them to capture richer spatial position information.

Metrics. Following VIEScore [14], we adopt three evaluation metrics: SC (Semantic Consistency), PQ (Perceptual Quality), and O (Overall Score). SC measures the consistency between the edited result and the given editing instruction, without considering image quality; PQ evaluates the edited image, including the consistency of the edited object, the stability of unedited regions, and whether the overall image appears realistic and free of artifacts. The overall score O is computed based on SC and PQ, defined as $O = (SC \times PQ)^{\frac{1}{2}}$. In our experiments, both GPT-5 [25] and

Table 1. **Quantitative comparison** on SpatialBench. GPT-SC, GPT-PQ, and GPT-O refer to the metrics evaluated by GPT-5, while Qwen-SC, Qwen-PQ, and Qwen-O refer to the metrics evaluated by Qwen3-VL-32B-Instruct. Scores range from 0 to 1. ↑: higher is better.

Method	GPT-SC↑	GPT-PQ↑	GPT-O↑	Qwen-SC↑	Qwen-PQ↑	Qwen-O↑
Flux-Kontext [2]	0.292	0.848	0.498	0.261	0.767	0.447
OmniGen2 [40]	0.301	0.661	0.446	0.305	0.687	0.458
Step1X-Edit [21]	0.484	0.785	0.616	0.492	0.691	0.583
BAGEL [47]	0.368	0.709	0.511	0.390	0.642	0.500
Qwen-Image-Editing [39]	0.666	0.882	0.766	0.632	0.813	0.717
SpatialDiff (Ours)	**0.803**	**0.886**	**0.843**	**0.778**	**0.838**	**0.807**

Qwen3-VL-32B-Instruct [33] were used to score the results according to these criteria.

5.2. Quantitative Comparison

As shown in Table 1, SpatialDiff achieves the best overall performance across all quantitative metrics for spatial movement task. In terms of Semantic Consistency (SC), it attains a GPT-SC score of 0.803 and a Qwen-SC score of 0.778, substantially outperforming existing methods. This result indicates that SpatialDiff precisely interprets and executes user instructions for spatial movement. For Perceptual Quality (PQ), both GPT-PQ (0.886) and Qwen-PQ (0.838) rank highest, reflecting strong preservation of global object consistency and natural visual appearance. Consequently, the overall score (O) also reaches the top (GPT-O: 0.843, Qwen-O: 0.807), confirming that SpatialDiff achieves a balanced improvement in both semantic accuracy and visual fidelity. In contrast, Flux-Kontext achieved artificially high PQ scores due to minimal image modifications, but their poor SC scores reveal the underlying failure in editing. While Qwen-Image-Editing achieves high image quality thanks to its powerful pretraining, its ability to follow instructions remains relatively limited (GPT-SC: 0.666, Qwen-SC: 0.632). Other methods (e.g., OmniGen2, Step1X-Edit, and BAGEL) struggle to balance editing fidelity and object consistency, resulting in suboptimal performance in both aspects. Although GPT and Qwen differ in scoring scales, their relative rankings of the methods are highly consistent, further validating the reliability of the results. In summary, SpatialDiff achieves state-of-the-art performance in semantic accuracy and visual fidelity.

5.3. Qualitative Comparison

Figure 4 presents qualitative results that highlight the precision and stability of our method in performing spatial movement. In line 1, our approach successfully relocates the table to the left side of the blue sofa while maintaining its appearance and consistency with the original scene. Our method reasoned in 3D, interpreting "left" relative to the sofa rather than the camera view. Other methods fail to achieve the correct spatial relocation, often introducing structural distortion or inconsistent object shapes.

Regarding scene coherence and background integrity, our method achieves superior global coordination. For line 3, when the cat is moved between the rabbit and the dog, background textures such as grass and shadows remain intact, and spatial relationships among objects are preserved naturally. Unedited regions also remain visually consistent, without redundant texture regeneration or boundary blurring. In comparison, OmniGen2 produces noticeable object deformation, Step1X-Edit merges foreground and background content incorrectly, BAGEL leaves ghost artifacts of the original cat, and Qwen-Image-Editing, although placing the cat approximately correctly, removes other unedited objects from the scene.

Our method also demonstrates strong 3D spatial understanding and depth consistency. In line 5, it correctly captures front–back relationships and depth ordering, naturally positioning the sculpture in front of the door. Other approaches such as Flux-Kontext and Qwen-Image-Editing appear reasonable in 2D projection but fail to maintain depth coherence, leading to weak spatial integration between objects and the door. Overall, qualitative comparisons show that our method excels in superior instruction following, global consistency, and spatial reasoning.

5.4. Ablations

Table 2 presents the ablation quantitative results on SpatialBench, systematically analyzing the contribution of each component in SpatialDiff.

Implicit 3D Spatial Modeling (ISM). For a fair comparison, we fine-tune the baseline model via LoRA on the movement dataset (Model A). This results in a noticeable improvement in SC (0.236 → 0.398 on GPT-SC) but a slight decrease in PQ (0.831 → 0.795). This indicates that while fine-tuning brings a limited improvement in the model's responsiveness to spatial instructions, it may also introduce unnatural distortions, highlighting the trade-off between instruction adherence and image quality. The visual observation ("parrot floating in mid-air") corroborates this quantitative evidence. Building upon this, introducing implicit 3D knowledge via a 3D foundation model (Model B) further boosts SC (0.398 → 0.518) and O (0.563 → 0.620), demonstrating that 3D-aware features substantially strengthen spa-

Table 2. **Ablation study** on SpatialBench. **FT** denotes fine-tuning the attention modules of DiT using LoRA. **ISM** (Implicit 3D Spatial Modeling) represents introducing 3D tokens from the input image through a 3D foundation model, and aligning them to the DiT feature space. **EDS** (Explicit Depth Supervision) indicates applying supervision on 3D tokens in the pixel space, while **LDS** (Latent Depth Supervision) applies supervision in the VAE latent space. GPT-SC, GPT-PQ, and GPT-O refer to the metrics evaluated by GPT-5, while Qwen-SC, Qwen-PQ, and Qwen-O refer to the metrics evaluated by Qwen3-VL-32B-Instruct. Scores range from 0 to 1. ↑: higher is better.

Method	FT	ISM	EDS	LDS	GPT-SC↑	GPT-PQ↑	GPT-O↑	Qwen-SC↑	Qwen-PQ↑	Qwen-O↑
Baseline					0.236	0.831	0.443	0.310	0.723	0.473
Model A	✓				0.398	0.795	0.563	0.427	0.739	0.562
Model B	✓	✓			0.518	0.743	0.620	0.513	0.684	0.592
Model C	✓	✓	✓		0.566	0.801	0.673	0.527	0.745	0.627
SpatialDiff	✓	✓		✓	**0.804**	**0.871**	**0.837**	**0.796**	**0.835**	**0.815**

tial reasoning and object localization.

Explicit Depth Supervision (EDS). Applying pixel-space depth supervision (Model C) enforces stronger spatial constraints, slightly improving O (0.620 → 0.673 on GPT-O; 0.592 → 0.627 on Qwen-O). Although this helps align spatial understanding, it causes over-constrained learning that affects non-edited regions (e.g., line 2 in Figure 5, the pillow on sofa should not be removed). Hence, EDS improves structural precision but at the expense of visual stability.

Latent Depth Supervision (LDS). When supervision is applied in the latent space (SpatialDiff), the model achieves the best overall balance, with large gains in all metrics (GPT-SC: 0.804, GPT-PQ: 0.871, GPT-O: 0.837). This indicates that latent-level guidance provides smooth, semantically coherent constraints that preserve global scene consistency and high perceptual quality simultaneously.

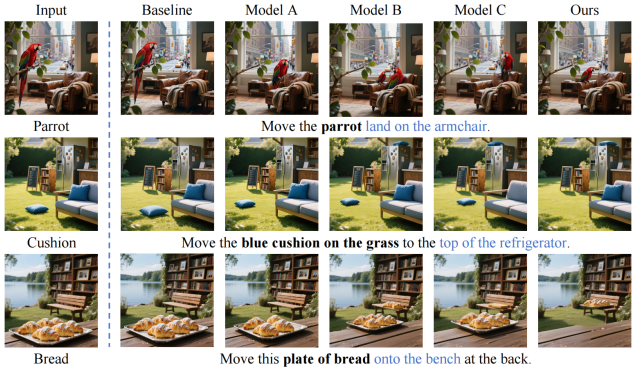


Figure 5. **Ablation study**. Model A denotes LoRA fine-tuning applied on the Baseline using the movement dataset; Model B denotes the introduction of Implicit 3D Spatial Modeling only; Model C denotes the incorporation of both Implicit 3D Spatial Modeling and Explicit Depth Supervision.

5.5. User Study

We conducted a user study with 35 participants, asking them to rate each generated image in terms of SC and PQ. Each participant evaluated ten results per method, resulting in 4200 votes. The results are reported in Table 3. Our method consistently achieves higher human-evaluated scores across all three metrics, H-SC, H-PQ, and H-O (overall performance). These results are generally consistent with the quantitative comparisons among the methods. The user study details are provided in supplementary material.

Table 3. **User Study**. H-SC refers to the assessment of instruction-following ability by humans, H-PQ refers to the evaluation of image quality by humans, and H-O represents the combined metric of both. Scores range from 1 to 5. ↑: higher is better.

Method	H-SC↑	H-PQ↑	H-O↑
Flux-Kontext [2]	1.874	3.433	2.536
OmniGen2 [40]	2.258	2.692	2.465
Step1X-Edit [21]	2.271	2.616	2.437
BAGEL [47]	2.759	3.351	3.041
Qwen-Image-Editing [39]	3.513	3.611	3.562
SpatialDiff (Ours)	**4.711**	**4.720**	**4.715**

6. Conclusion

In this work, we introduced SpatialDiff, a novel instruction-based method designed to enable precise and consistent object movement in complex scenes. Unlike previous 2D editing methods that rely solely on planar priors, SpatialDiff incorporates 3D spatial awareness into the editing process through two key designs: Implicit 3D Spatial Modeling and Global Spatial Supervision. Implicit 3D Spatial Modeling integrates 3D knowledge into the latent space to enhance the model's understanding of spatial information, while Global Spatial Supervision provides latent-level constraints on the modeled features, smoothly guiding consistent object movement during training. Experiments demonstrate that SpatialDiff achieves superior spatial reasoning and perceptual fidelity, producing geometrically coherent results even with occlusions and multi-depth structures. We believe this study opens new avenues for integrating 3D understanding into diffusion models for spatially-perceptive image editing.

Acknowledgments

This paper is supported by the National Natural Science Foundation of China (No. 62322608), and is also sponsored by CCF-Kuaishou Large Model Explorer Fund (No. CF-KuaiShou 2024007).

References

[1] Hadi Alzayer, Zhihao Xia, Xuaner Zhang, Eli Shechtman, Jia-Bin Huang, and Michael Gharbi. Magic fixup: Streamlining photo editing by watching dynamic videos. *ACM Transactions on Graphics*, 44(5):1–25, 2025. 1

[2] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025. 2, 5, 7, 8, 3

[3] Shariq Farooq Bhat, Niloy Mitra, and Peter Wonka. Loosecontrol: Lifting controlnet for generalized depth conditioning. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. 1, 3

[5] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22560–22570, 2023. 1

[6] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6593–6602, 2024. 1

[7] En Ci, Shanyan Guan, Yanhao Ge, Yilin Zhang, Wei Li, Zhenyu Zhang, Jian Yang, and Ying Tai. Describe, don't dictate: Semantic image editing with natural language intent. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19185–19194, 2025. 1

[8] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 1

[9] Kunyu Feng, Yue Ma, Bingyuan Wang, Chenyang Qi, Haozhe Chen, Qifeng Chen, and Zeyu Wang. Dit4edit: Diffusion transformer for image editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2969–2977, 2025. 1

[10] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102*, 2023. 2, 3

[11] Junjia Huang, Pengxiang Yan, Jiyang Liu, Jie Wu, Zhao Wang, Yitong Wang, Liang Lin, and Guanbin Li. Dreamfuse: Adaptive image fusion with diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17292–17301, 2025. 1

[12] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8362–8371, 2024. 1, 2, 3

[13] Boseong Jeon, Junghyuk Lee, Jimin Park, Kwanyoung Kim, Jingi Jung, Sangwon Lee, and Hyunbo Shim. Crimedit: Controllable editing for counterfactual object removal, insertion, and movement. *arXiv preprint arXiv:2509.23708*, 2025. 2

[14] Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12268–12290, 2024. 6

[15] Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19721–19730, 2025. 2, 3

[16] Nupur Kumari, Grace Su, Richard Zhang, Taesung Park, Eli Shechtman, and Jun-Yan Zhu. Customizing text-to-image diffusion with object viewpoint control. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–13, 2024. 3

[17] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 1

[18] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 2, 3, 5

[19] Yaowei Li, Lingen Li, Zhaoyang Zhang, Xiaoyu Li, Guangzhi Wang, Hongxiang Li, Xiaodong Cun, Ying Shan, and Yuexian Zou. Blobctrl: A unified and flexible framework for element-level image generation and editing. *arXiv preprint arXiv:2503.13434*, 2025. 1

[20] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 3

[21] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025. 2, 3, 5, 7, 8

[22] Léopold Maillard, Tom Durand, Adrien Ramanana Rahary, and Maks Ovsjanikov. Laconic: A 3d layout adapter for controllable image creation. In *Proceedings of the IEEE/CVF*

International Conference on Computer Vision, pages 18046–18057, 2025. 2, 3

[23] Oscar Michel, Anand Bhattad, Eli VanderBilt, Ranjay Krishna, Aniruddha Kembhavi, and Tanmay Gupta. Object 3dit: Language-guided 3d-aware image editing. *Advances in Neural Information Processing Systems*, 36:3497–3516, 2023. 5

[24] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Diffeditor: Boosting accuracy and flexibility on diffusion-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8488–8497, 2024. 3

[25] OpenAI. Introducing gpt-5. https://openai.com/research/introducing-gpt-5, 2025. Accessed November 2025. 6

[26] Karran Pandey, Paul Guerrero, Matheus Gadelha, Yannick Hold-Geoffroy, Karan Singh, and Niloy J Mitra. Diffusion handles enabling 3d edits for diffusion models by lifting activations to 3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7695–7704, 2024. 2, 3

[27] Or Patashnik, Rinon Gal, Daniel Cohen-Or, Jun-Yan Zhu, and Fernando De la Torre. Consolidating attention features for multi-view image editing. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024. 3

[28] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 1

[29] Tianhao Qi, Shancheng Fang, Yanze Wu, Hongtao Xie, Jiawei Liu, Lang Chen, Qian He, and Yongdong Zhang. Deadiff: An efficient stylization diffusion model with disentangled representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8693–8702, 2024. 1

[30] Jiawei Ren, Mengmeng Xu, Jui-Chieh Wu, Ziwei Liu, Tao Xiang, and Antoine Toisoul. Move anything with layered scene diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6380–6389, 2024. 3

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[32] Qianqian Sun, Jixiang Luo, Dell Zhang, and Xuelong Li. Smartfreeedit: Mask-free spatial-aware image editing with complex instruction understanding. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 8244–8252, 2025. 2

[33] Qwen Team. Qwen3 technical report, 2025. 5, 7

[34] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pages 439–457. Springer, 2024. 3

[35] Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. *arXiv preprint arXiv:2411.04746*, 2024. 1

[36] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 3, 5

[37] Ruicheng Wang, Jianfeng Xiang, Jiaolong Yang, and Xin Tong. Diffusion models are geometry critics: Single image 3d editing using pre-trained diffusion priors. In *European Conference on Computer Vision*, pages 441–458. Springer, 2024. 2, 3

[38] Ye Wang, Ruiqi Liu, Jiang Lin, Fei Liu, Zili Yi, Yilin Wang, and Rui Ma. Omnistyle: Filtering high quality style transfer data at scale. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7847–7856, 2025. 1

[39] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. 2, 3, 5, 7, 8

[40] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. 2, 5, 7, 8, 3

[41] Jay Zhangjie Wu, Xuanchi Ren, Tianchang Shen, Tianshi Cao, Kai He, Yifan Lu, Ruiyuan Gao, Enze Xie, Shiyi Lan, Jose M Alvarez, et al. Chronoedit: Towards temporal reasoning for image editing and world simulation. *arXiv preprint arXiv:2510.04290*, 2025. 2, 3

[42] Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18682–18692, 2025. 5

[43] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 5

[44] Jiraphon Yenphraphai, Xichen Pan, Sainan Liu, Daniele Panozzo, and Saining Xie. Image sculpting: Precise object editing with 3d geometry control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4241–4251, 2024. 2

[45] Xin Yu, Tianyu Wang, Soo Ye Kim, Paul Guerrero, Xi Chen, Qing Liu, Zhe Lin, and Xiaojuan Qi. Objectmover: Generative object movement with video prior. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17682–17691, 2025. 2

[46] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 1

[47] Xinjie Zhang, Jintao Guo, Shanshan Zhao, Minghao Fu, Lunhao Duan, Jiakui Hu, Yong Xien Chng, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, et al. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2505.02567*, 2025. 2, 3, 5, 7, 8

[48] Haozhe Zhao, Xiaojian Shawn Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems*, 37:3058–3093, 2024. 1

[49] Weizhi Zhong, Huan Yang, Zheng Liu, Huiguo He, Zijian He, Xuesong Niu, Di Zhang, and Guanbin Li. Mod-adapter: Tuning-free and versatile multi-concept personalization via modulation adapter. *arXiv preprint arXiv:2505.18612*, 2025.

[50] Jun Zhou, Jiahao Li, Zunnan Xu, Hanhui Li, Yiji Cheng, Fa-Ting Hong, Qin Lin, Qinglin Lu, and Xiaodan Liang. Fireedit: Fine-grained instruction-based image editing via region-aware vision language model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13093–13103, 2025. 1, 2, 3

[51] Hanshen Zhu, Zhen Zhu, Kaile Zhang, Yiming Gong, Yuliang Liu, and Xiang Bai. Training-free geometric image editing on diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19130–19140, 2025. 2, 3