GeoEdit: Geometry-Aware Object Editing via Dual-Branch Denoising

Yi He[1,*], Jiangming Wang[3,*], Xinyu Wang[1], Mark Fong[4], Songchun Zhang[5], Yuxuan Xue[6,‡], Hai-Tao Zheng[1,2,†], and Yue Ma[5,†]

[1] Shenzhen International Graduate School, Tsinghua University
[2] Pengcheng Laboratory, Shenzhen, China
[3] Sun Yat-sen University
[4] Peking University
[5] Hong Kong University of Science and Technology
[6] University of Tübingen

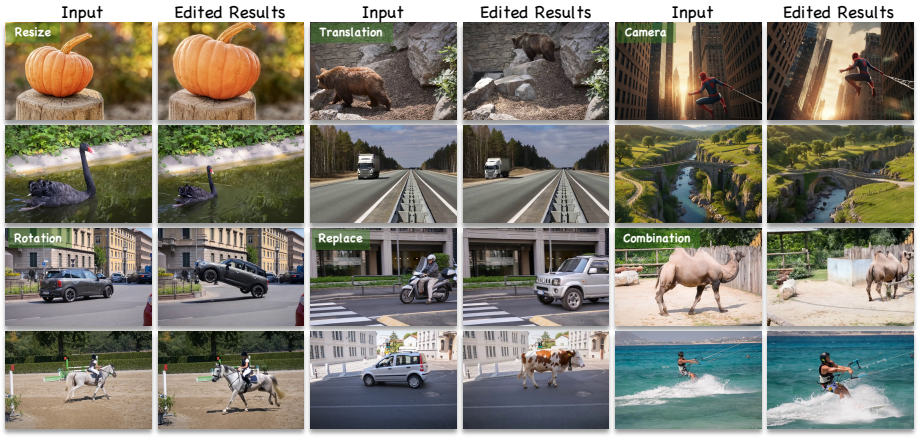

Fig. 1: Showcase of proposed GeoEdit. In this paper, we propose GeoEdit, a training-free pipeline that lifts editing into 3D for physically plausible object manipulation, without external 3D software or synthetic training data.

Abstract. Precisely manipulating objects in a single photograph (translation, rotation, scaling) while obeying 3D physical constraints remains unsolved for diffusion-based editors. Current 2D methods lack spatial awareness and produce perspective violations. Forcing structural proxies into the latent space also disrupts variance homogeneity, and the resulting self-attention leakage leads to ghosting and background blur. The core difficulty is asymmetric: the relocated object must follow a rigid geometry, yet the uncovered background needs freedom to synthesize plausible content. We present **GeoEdit**, a training-free *Lift-Manipulate-Render-Denoise* pipeline that satisfies both constraints. We decouple scene and object in 3D, align them through point correspondence, and

* Equal contribution.
‡ Project leader.
† Corresponding authors.

render a geometry-aligned proxy with a structural depth map. A Dual-Branch Denoising stage then refines this proxy: a video diffusion backbone preserves object identity, while 3D constraints are injected into the foreground within a narrow denoising window at matching noise variance (variance-homogeneous injection). The background denoises freely. Because the injected signal matches the native latent statistics, self-attention stays undisturbed. We also introduce GeoEditBench, a pose-aware benchmark covering object translation, object rotation, and camera movement with pose-aware evaluation metrics. Experiments confirm consistent gains in geometric accuracy, identity fidelity, and background quality, validated by automatic metrics and human studies. Code and data are publicly available at https://github.com/Heey731/GeoEdit.

Keywords: Image Editing · 3D-Aware Manipulation · Diffusion Models

1 Introduction

Image diffusion models [14, 17, 21–23, 47, 54–56] have advanced synthesis and editing, but one task remains open: precisely manipulating objects in 3D, translating, rotating, or scaling them within a photograph while respecting physical constraints. Such control matters for content creation and augmented reality, where edits must obey perspective, occlusion, and scene geometry.

Existing approaches, primarily mask-based inpainting [26], operate entirely in the 2D pixel plane and have three systematic weaknesses. First, without 3D spatial awareness, modifying object coordinates in the image domain violates perspective constraints and produces distorted geometry. Second, the strong generative prior of diffusion models tends to hallucinate remnants of the original object at its old position, a phenomenon called ghosting. Third, forcing structural proxies such as hard masks into the latent space introduces a distribution mismatch: abrupt interventions break the variance homogeneity that Diffusion Transformers [43] assume, and the resulting self-attention leakage causes background blur and structural collapse.

These failures share a common root: the edited foreground and the uncovered background impose asymmetric generative demands. The relocated object must rigidly follow a prescribed 3D geometry, while the exposed background, which now contains disoccluded holes and the original object's footprint, needs generative freedom to synthesize plausible content. No single denoising trajectory can satisfy both at once.

We introduce GeoEdit, a framework that resolves this conflict through a principled *Lift-Manipulate-Render-Denoise* pipeline. Inspired by decoupled 3D reconstruction approaches [5], our lifting phase independently recovers the global scene and a geometry-complete foreground object, then registers them through cross-space point matching into a unified coordinate frame. Users apply precise spatial manipulations, and the manipulated object is rendered back into a geometry-aligned proxy image alongside a structural depth map that provides rigid conditioning for generation.

To turn this coarse proxy into a photorealistic composite, we propose Dual-Branch Denoising. We repurpose a video diffusion backbone [60] whose temporal prior inherently preserves object identity; no identity-specific fine-tuning is needed. Motivated by the region-dependent scheduling of Time-to-Move [53], we inject the 3D proxy into the foreground region only within a carefully selected denoising window. The injection uses variance-homogeneous noise matching so that the modified latent is statistically indistinguishable from the native denoising path; this prevents self-attention leakage. Outside the window, the generative prior freely removes the original object and fills in the background. The result is an asymmetric treatment of foreground rigidity and background flexibility, without architectural changes or additional training.

The main contributions of this work are:

- A geometry-aware training-free editing framework based on a *Lift-Manipulate-Render-Denoise* pipeline that lifts editing into 3D for physically plausible object manipulation, without external 3D software [8] or synthetic training data [38].
- Dual-Branch Denoising, which uses a video diffusion backbone for identity preservation and injects 3D constraints via variance-homogeneous injection within a selective denoising window, resolving the asymmetric constraint between foreground rigidity and background flexibility.
- GeoEditBench, a benchmark covering object translation, object rotation, and camera movement with pose-aware evaluation metrics, along with consistent improvements over existing methods.

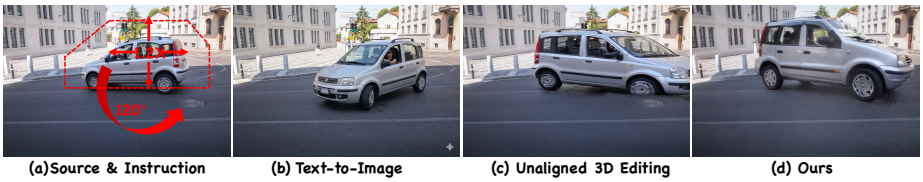

Fig. 2: Comparison with previous approaches on geometry-aware object manipulation. Given a source image and an instruction requiring a 120° rotation and 1.2× scaling, existing methods struggle to maintain geometric consistency. In contrast, our method faithfully follows the specified transformation while preserving object identity and producing coherent, realistic results.

2 Related Work

Image editing with diffusion models. GAN-based models [18, 20] first enabled controllable editing through latent manipulation [50] and text-driven synthesis [42]. Diffusion models [9, 17, 28–36, 47, 51, 67] have since become the dominant paradigm: InstructPix2Pix [4] supports text-guided editing, and RePaint [26]

enables mask-based inpainting. ControlNet [73] adds spatial conditions such as depth [46] and edges [66], while DragGAN [40] and DragDiffusion [52] allow point-based manipulation. These methods all operate in the 2D image plane and struggle with out-of-plane transformations that require 3D spatial reasoning. Similarly, recent 2D editing and insertion methods such as FateZero [44], Ctrl&Shift [48], and ObjectAdd [75] provide useful semantic or composition control but lack explicit physical constraints for prescribed 3D transformations.

Geometry-aware object editing. Single-image lifting methods [12, 15, 16, 24, 25, 27, 39, 61, 62, 69, 70, 76, 77] synthesize novel views but often lack appearance consistency when composited into real scenes. Dataset-driven approaches [7, 11] suffer from synthetic-to-real domain gaps [64, 68]. Among diffusion-based works, GeoDiffuser [49] and Diffusion Handles [41] inject 3D conditions, while Object-3DIT [38] uses language guidance but relies on synthetic data. ObjectMover [72] uses video priors for translation, and BlenderFusion [8] requires external 3D software. Recently, 3DitScene [74] lifts pixels to 3D Gaussians for Score Distillation Sampling (SDS) editing. While semantic alignment is competitive, SDS incurs high computational costs and yields weaker geometric correctness. In contrast, our training-free pipeline uses monocular lifting to support full rigid transformations (rotation, scaling) without external software or synthetic data.

Denoising strategies for structured editing. Injecting structural constraints into pre-trained diffusion models without compromising generation quality remains a core challenge. SDEdit [37] balances fidelity and realism via timestep-based initialization, while Blended Diffusion [2] and DiffEdit [10] rely on masked blending, often causing boundary artifacts or distribution mismatch. More recently, Time-to-Move [53] explores region-dependent scheduling for video generation. Our Dual-Branch Denoising resolves the asymmetric tension between rigid foreground geometry and generative background freedom via variance-homogeneous injection within a selective denoising window, which preserves latent statistics and effectively suppresses self-attention leakage.

3 Method

We present a geometry-aware framework for precise object manipulation. As shown in Fig. 3, our pipeline follows a 2D→3D→2D workflow. While the 3D lifting and rendering modules leverage existing techniques, the core contribution of GeoEdit is a dual-branch denoising architecture with variance-homogeneous injection. This mechanism mitigates attention leakage by aligning injected noise statistics with the pre-trained diffusion backbone, enabling rigid foreground transformation alongside free-form background synthesis. The source image is first lifted into a 3D representation for user-defined spatial transformations, rendered into a 2D proxy (I_{proxy}) with a depth map, and finally converted into a photorealistic composite via our dual-branch denoiser (Sec. 3.2, 3.3).

Problem Formulation. Given a source image $I_{\text{src}} \in \mathbb{R}^{H\times W\times 3}$ and a user-specified 3D transformation $\mathcal{T} = \{\mathbf{R}, \mathbf{t}, s\}$ on a target object, the framework produces geometrically aligned conditioning signals: a geometry-aligned proxy

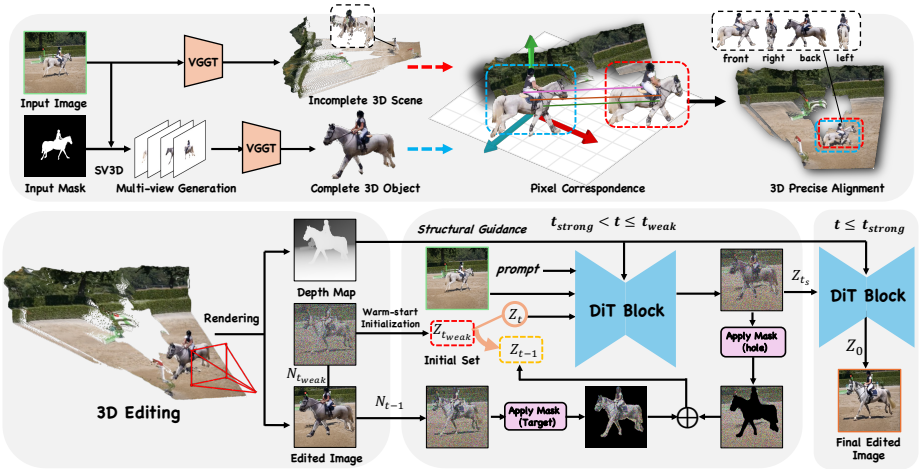

Fig. 3: Overview of the proposed framework. Top: Decoupled 3D reconstruction and precise alignment pipeline. Bottom: Dual-branch denoising architecture featuring warm-start initialization and variance-homogeneous injection.

$I_{\text{proxy}} \in \mathbb{R}^{H\times W\times 3}$, a structural depth map $D_{\text{rep}} \in \mathbb{R}^{H\times W}$, and a spatial mask $M \in \{0,1\}^{H\times W}$ obtained by projecting the transformed object silhouette back into the image plane. The goal is to generate a realistic composite $x_0 \in \mathbb{R}^{H\times W\times 3}$ that strictly respects $\mathcal{T}$ while preserving the unedited background of I_{src}.

3.1 Preliminaries: Latent Diffusion Models

Latent Diffusion Models (LDMs) [17,47] operate in a compressed latent space to reduce computational cost while maintaining generation quality. Given an RGB image $x \in \mathbb{R}^{H\times W\times 3}$, a pre-trained encoder maps it to a latent representation $z_0 = \mathcal{E}(x)$. The forward diffusion process gradually adds Gaussian noise to z_0 over T steps. At timestep $t \in [1, T]$, the noisy latent z_t is:

$$z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{1}$$

where $\bar{\alpha}_t$ follows a predefined noise schedule. During the reverse process, a network ϵ_θ, parameterized as a UNet or Diffusion Transformer (DiT), is trained to predict the added noise ϵ given t and optional conditioning signals c:

$$L_{LDM} = \mathbb{E}_{z_0,\epsilon\sim\mathcal{N}(0,\mathbf{I}),t,c}\left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2\right] \tag{2}$$

Once sampling is complete, a decoder maps the denoised latent back to pixel space: $x' = \mathcal{D}(z_0)$. Our method does not fine-tune any model weights; it intervenes in the reverse sampling process to inject 3D geometric constraints into a pre-trained DiT.

3.2 Decoupled 3D Reconstruction and Canonical Alignment

To overcome the occlusion ambiguity inherent in monocular lifting, we reconstruct the global scene and the target object independently.

Decoupled 3D Reconstruction. As shown in Fig. 3, we first use a monocular depth estimator [63] to lift the source image I_{src} into an incomplete 3D scene. The spatial mask M segments this scene into a background $\mathcal{P}_{bg}$ and a visible foreground anchor $\mathcal{P}_{fg}^{\text{vis}}$. To enable precise manipulation, the occluded surfaces of the target object must be recovered. We apply a multi-view diffusion model [59] to the masked foreground to synthesize novel views, constructing a geometry-complete object point cloud $\mathcal{P}_{fg}^{\text{comp}}$ in an isolated canonical space.

Correspondence-Aware Alignment. To perform physically plausible edits, the isolated object must be registered back into the global scene. Both $\mathcal{P}_{fg}^{\text{comp}}$ and $\mathcal{P}_{fg}^{\text{vis}}$ originate from I_{src}, so points at the same pixel coordinate $\mathbf{u}$ correspond to the same physical surface point [5]. We use this to extract dense 3D–3D correspondences:

$$\mathcal{C} = \left\{ \left(\mathcal{P}_{fg}^{\text{comp}}(\mathbf{u}),\, \mathcal{P}_{fg}^{\text{vis}}(\mathbf{u}) \right) \mid M(\mathbf{u}) = 1 \right\} \tag{3}$$

From the correspondence set $\mathcal{C}$, we compute the optimal similarity transformation (rotation $\mathbf{R}_{\text{align}}$, translation $\mathbf{t}_{\text{align}}$, scale s_{align}) by minimizing the registration error:

$$\underset{\mathbf{R}_{\text{align}},\, \mathbf{t}_{\text{align}},\, s_{\text{align}}}{\arg\min} \sum_{(\mathbf{p},\mathbf{q}) \in \mathcal{C}} \| s_{\text{align}}\, \mathbf{R}_{\text{align}}\, \mathbf{p} + \mathbf{t}_{\text{align}} - \mathbf{q} \|_2^2 \tag{4}$$

where $\mathbf{p} = \mathcal{P}_{fg}^{\text{comp}}(\mathbf{u})$ and $\mathbf{q} = \mathcal{P}_{fg}^{\text{vis}}(\mathbf{u})$ denote the matched 3D points. This aligns $\mathcal{P}_{fg}^{\text{comp}}$ into the global scene space, establishing a unified coordinate frame in which the user can precisely apply the desired manipulation $\mathcal{T}$.

Proxy Rendering. To render the final conditioning signals, the original visible foreground is removed, leaving an exposed background hole in I_{src}. Instead of relying on a black-box heuristic, this hole is efficiently filled using the Telea inpainting algorithm [58] based on the fast marching method, providing a coarse but structurally continuous global color layout. The aligned and manipulated 3D object is then re-projected over this completed background to form the geometry-aligned proxy I_{proxy} and the structural depth map D_{rep}. While the background texture is coarse at this stage, the downstream denoising uses it as a structural baseline and synthesizes realistic high-frequency details.

3.3 Dual-Branch Denoising and Variance Injection

After the 3D editing and rendering process, the geometry-aligned proxy I_{proxy} has correct 3D structure but coarse textures and exposed background holes. Adapting this proxy via standard single-timestep initialization (SDEdit [37]) presents a trade-off between structural preservation and semantic realism. As

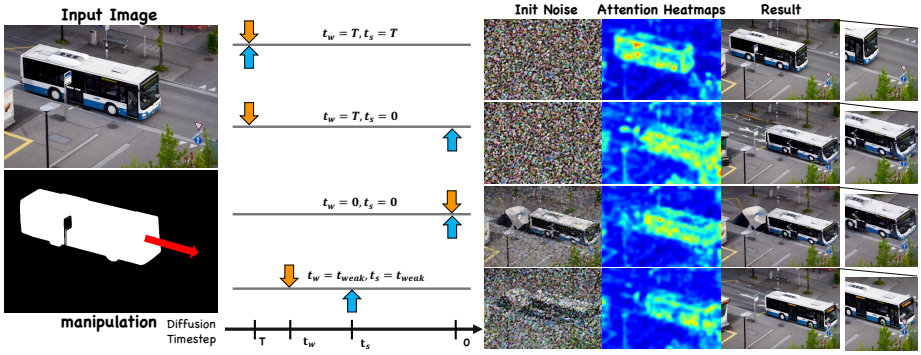

Fig. 4: Visualizing the generative trade-off. Different configurations of initialization (t_{weak}) and injection (t_{strong}) timesteps dictate whether the model leans towards preserving the rigid object skeleton or hallucinating semantic background details.

the attention heatmaps in Fig. 4 illustrate, a large initialization timestep affords the generative prior enough freedom to hallucinate realistic backgrounds, but inevitably degrades the prescribed 3D object skeleton, causing severe geometric drift. Conversely, a small timestep forces strict adherence to the proxy but retains coarse, unrealistic background artifacts.

We argue that the manipulated foreground and the unedited background require asymmetric generative constraints: the target object demands rigid adherence to the 3D spatial signal, whereas the background requires generative flexibility. Furthermore, standard DiTs assume a globally homogeneous noise distribution. Directly forcing mismatched latents into specific regions disrupts this homogeneity, causing self-attention leakage and global blurring. To reconcile these constraints, we introduce a training-free *Dual-Branch Denoising* strategy (Fig. 3, bottom).

Video Diffusion Backbone. We build on a depth-conditioned video diffusion model Φ_{VDM} [19, 60] as our generation backbone. The source image I_{src} is the appearance reference, D_{rep} provides structural depth conditioning through ControlNet-style control blocks [73] that inject geometric features into the transformer, and a text prompt c guides semantic context:

$$x_0 = \Phi_{\text{VDM}}\left(I_{\text{src}}, D_{\text{rep}}, c\right)\big|_{\text{last frame}} \tag{5}$$

Depth maps act as a geometric scaffold that anchors the object's skeleton against degradation during high-noise phases. To repurpose this video model for single-image editing, we construct a short target sequence by repeating the proxy as identical frames; the temporal self-attention of Φ_{VDM} then enforces cross-frame consistency, which naturally preserves the object's identity without any identity-specific fine-tuning or adapters. The last generated frame is taken as the output.

Warm-Start Initialization. We avoid the semantic ambiguity of pure noise ($t = T$) by initializing the reverse process at an intermediate timestep t_{weak}. We

Algorithm 1 Dual-Branch Denoising

Require: Source image I_{src}, proxy I_{proxy}, depth D_{rep}, mask M, window bounds t_{weak}, t_{strong}, video backbone Φ_{VDM}
Ensure: Edited composite x_0
$z_{\text{proxy}} \leftarrow \mathcal{E}(I_{\text{proxy}})$ {Encode proxy to latent space}
$\epsilon' \sim \mathcal{N}(0, \mathbf{I})$ {Sample fixed noise (reused across all steps)}
$z_{t_{\text{weak}}} \leftarrow \sqrt{\bar{\alpha}_{t_{\text{weak}}}}\, z_{\text{proxy}} + \sqrt{1-\bar{\alpha}_{t_{\text{weak}}}}\, \epsilon'$ {Warm-start initialization}
for $t = t_{\text{weak}}, \ldots, 1$ **do**
  $z_{t-1}^{\text{pred}} \leftarrow \text{Denoise}(\Phi_{\text{VDM}}, z_t, t, I_{\text{src}}, D_{\text{rep}})$ {Backbone step}
  if $t_{\text{strong}} < t \leq t_{\text{weak}}$ **then**
    $z_{t-1}^{\text{ref}} \leftarrow \sqrt{\bar{\alpha}_{t-1}}\, z_{\text{proxy}} + \sqrt{1-\bar{\alpha}_{t-1}}\, \epsilon'$ {Variance-matched proxy}
    $z_{t-1} \leftarrow M \odot z_{t-1}^{\text{ref}} + (1-M) \odot z_{t-1}^{\text{pred}}$ {Replace foreground only}
  else
    $z_{t-1} \leftarrow z_{t-1}^{\text{pred}}$ {Outside window: free running}
  end if
end for
$x_0 \leftarrow \mathcal{D}(z_0)$ {Decode to pixel space}

encode I_{proxy} and perturb it to t_{weak}:

$$z_{t_{\text{weak}}} = \sqrt{\bar{\alpha}_{t_{\text{weak}}}}\, \mathcal{E}(I_{\text{proxy}}) + \sqrt{1-\bar{\alpha}_{t_{\text{weak}}}}\, \epsilon', \quad \epsilon' \sim \mathcal{N}(0, \mathbf{I}) \tag{6}$$

This establishes a global color layout and structural baseline, so the model retains the scene's overall identity from the start of the denoising trajectory.

Variance-Homogeneous Injection. To enforce asymmetric constraints, we perform selective latent replacement within a denoising window $t_{\text{strong}} < t \leq t_{\text{weak}}$. At each step t within this window, instead of raw feature replacement, we synchronize the masked object region with the forward-noised proxy [53]. Building on the masked blending of Blended Diffusion [2] but strictly maintaining variance homogeneity, we formulate the injection as:

$$\tilde{z}_{t-1} = M \odot \left(\sqrt{\bar{\alpha}_{t-1}}\, \mathcal{E}(I_{\text{proxy}}) + \sqrt{1-\bar{\alpha}_{t-1}}\, \epsilon' \right) + (1-M) \odot z_{t-1}^{\text{pred}} \tag{7}$$

where z_{t-1}^{pred} is the latent predicted by the Φ_{VDM} backbone, and $\epsilon' \sim \mathcal{N}(0, \mathbf{I})$ is a fixed Gaussian noise tensor sampled once at the start. The foreground thus tracks the proxy's 3D geometry, while the background z_{t-1}^{pred} evolves under the generative prior. Because the injected signal has exactly the variance that the noise schedule prescribes at timestep $t-1$, it is statistically indistinguishable from the native denoising path, and the self-attention mechanism sees a spatially homogeneous distribution. To quantitatively validate this, we introduce the Attention Leakage Ratio (ALR) to measure the fraction of background-query attention mass assigned to foreground proxy tokens, defined as ALR=$\text{mean}_{l,h} A^{l,h}_{\mathcal{B}\to\mathcal{F}} / A^{l,h}_{\mathcal{B}\to *}$. Our analysis confirms that this variance-matched injection effectively suppresses self-attention leakage, notably reducing the ALR from 8.4% to 6.4% at the peak leakage step ($t = 46$). Fixing ϵ' across all steps maintains structural consistency

of the injected proxy. The complete dual-branch denoising procedure is summarized in Algorithm 1.

Global Harmonization. Once the denoising crosses $t \leq t_{\text{strong}}$, the mask injection loop terminates. The latent evolves freely under the pre-trained generative prior. During this phase, the model harmonizes mask boundaries and synthesizes high-frequency textures to fill disoccluded holes. The final decoded frame is the output composite.

4 Experiments

4.1 Implementation Details

Our framework operates as a fully training-free pipeline, strategically orchestrating off-the-shelf pre-trained models across all intermediate stages. Specifically, we adopt VGGT [63] for dense 3D point-map reconstruction from both the source image and the synthesized multi-view observations. To achieve canonical object completion, SV3D [59] is utilized to generate geometrically consistent novel views of the isolated foreground. The terminal dual-branch denoising phase is powered by Wan2.2-VACE [19, 60], which serves as our depth-conditioned generative prior.

All experiments are conducted on a single NVIDIA A800 GPU at a standardized resolution of 720×480 pixels. For variance-homogeneous injection, we use a diffusion schedule of $T = 50$ timesteps, setting the warm-start threshold to $t_{\text{weak}} = 47$ and ending mask injection at $t_{\text{strong}} = 40$. This empirically determined window achieves the best trade-off between geometric fidelity and background harmonization. Runtime analysis shows that the 3D extraction stage requires approximately 2 minutes and 19 GB of VRAM, whereas the denoising stage takes about 16 minutes and 44 GB of VRAM.

4.2 GeoEditBench

To systematically evaluate geometry-aware image editing, we introduce GeoEditBench, a benchmark of 200 image pairs collected from diverse real-world scenarios. We curate the dataset by removing samples with ambiguous geometry, severe rendering artifacts, or inconsistent illumination. The benchmark is divided into three categories according to the primary spatial transformation: 80 **object translation** examples, 80 **object rotation** examples, and 40 **camera movement** examples. This category-aware design enables both overall and per-category evaluation of spatial editing performance under diverse geometric manipulations.

4.3 Comparison with Baselines

Quantitative comparison. Table 1 reports the quantitative zero-shot evaluation of our method against several recent baselines on GeoEditBench, evaluated

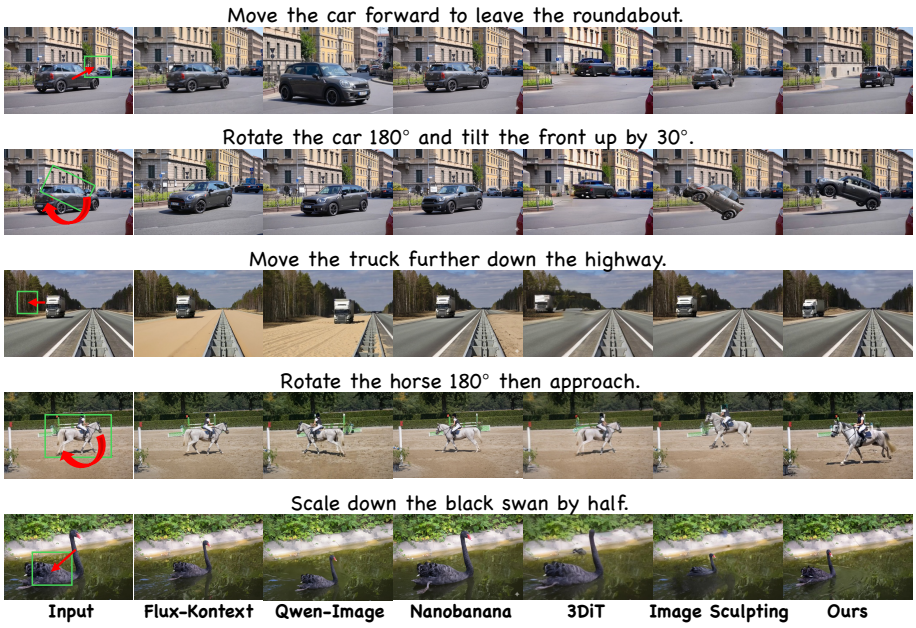

Fig. 5: Qualitative comparison of different methods on object manipulation tasks. Our model achieves superior performance compared to state-of-the-art methods in background preservation and geometric consistency.

on the same pre-defined fixed 50-pair stratified subset used throughout Tabs. 1-3. Following prior work, we evaluate performance across three dimensions: (i) reconstruction fidelity using PSNR; (ii) object identity preservation using DINO [6] and CLIP similarity [45]; and (iii) perceptual discrepancy using LPIPS and DreamSim [13]. To assess the physical correctness of 3D manipulations, we further introduce PoseMap IoU and Object IoU. PoseMap IoU compares predicted and target pose maps, while Object IoU measures silhouette alignment with the transformed target mask; both rely on external estimators independent of GeoEdit. Overall, our method achieves the strongest performance on most metrics, obtaining the highest PSNR (23.499) and DINO (0.961) alongside the lowest LPIPS (0.114) and DreamSim (0.027). On the geometry-aware metrics, GeoEdit further achieves 94.9% PoseMap IoU and 57.9% Object IoU, demonstrating precise geometric control. Although NanoBanana attains a slightly higher CLIP score (0.976 vs. 0.952), our method provides the strongest overall balance between geometric correctness, identity preservation, perceptual quality, and background fidelity. Additional evaluations (3DEdit-Bench, per-category results, and confidence intervals) are provided in the Supplementary Material.

Qualitative comparison. Figure 5 presents a qualitative comparison between our method and several state-of-the-art approaches, including Image Sculpting [71], NanoBanana [1], Flux-Kontext [3], Qwen-Image-Edit [65], and 3DiT [38].

Table 1: Comparison with state-of-the-art object editing methods on the fixed 50-pair stratified subset of GeoEditBench. Preference scores (Pref. ↑) indicate VLM (AI) and human ratings. Geometry-aware metrics explicitly evaluate 3D transformation correctness. Red and Blue denote the best and second best results.

Method	Quantitative metrics							Pref. ↑	
	CLIP ↑	PSNR ↑	DINO ↑	LPIPS ↓	DreamSim ↓	PoseMap IoU ↑	Object IoU ↑	AI	Human
Qwen-Image [65]	0.885	14.013	0.685	0.431	0.185	66.7%	8.8%	2.010	2.667
NanoBanana [1]	**0.976**	19.164	**0.948**	**0.118**	**0.031**	75.0%	25.2%	3.085	2.286
3DiT [38]	0.941	21.532	0.771	0.371	0.084	60.0%	33.5%	1.300	1.191
Flux-Kontext [3]	0.927	20.250	0.841	0.261	0.112	66.7%	**51.3%**	2.563	2.714
Image Sculpting [71]	0.939	**22.101**	0.802	0.147	0.104	**80.0%**	28.2%	2.215	2.619
Ours	**0.952**	**23.499**	**0.961**	**0.114**	**0.027**	**94.9%**	**57.9%**	**4.125**	**4.810**

3DiT exhibits limited generalization to real-world data, likely due to its reliance on specific training datasets. Image Sculpting demonstrates relatively strong object manipulation capability; however, it tends to lose fine-grained details, which affects overall visual fidelity. Flux-Kontext and Qwen-Image-Edit show some limitations in maintaining background consistency, leading to noticeable inconsistencies in certain surrounding regions after editing. NanoBanana preserves background content comparatively well, but its ability to perform precise object manipulation appears less robust in complex spatial adjustments. In contrast, our method achieves more reliable camera pose control and object relocation while maintaining consistent background appearance, resulting in visually coherent and geometrically aligned outputs.

Table 2: Quantitative Ablation on Timestep Thresholds. We evaluate the trade-off between geometric structure preservation and semantic appearance generation. Our configured interval ($t_w = 47, t_s = 40$) achieves the optimal balance across all metrics.

Configuration (t_w, t_s)	PSNR ↑	DINO ↑	CLIP ↑	DreamSim ↓
Pure Prior (50, 50)	18.479	0.894	0.921	0.057
SDEdit Init (47, 47)	20.724	0.932	0.921	0.044
Low Noise Init (40, 40)	21.822	0.942	0.906	0.060
Full Injection (50, 0)	20.012	0.860	0.871	0.050
Pure Proxy (1, 1)	19.465	0.812	0.806	0.084
Ours Optimal (47, 40)	**23.499**	**0.961**	**0.952**	**0.027**

Human and AI Preference Evaluation. Given the inherently subjective nature of generative editing, we conduct comprehensive qualitative assessments utilizing both human participants and an advanced Vision-Language Model (VLM). We evaluate the generated results based on three core criteria: (1) *Object Identity Preservation*, assessing whether the manipulated target strictly retains its original appearance; (2) *Instruction Fidelity*, measuring how accurately the visual edit reflects the provided spatial command; and (3) *Background Preservation*, evaluating the consistency and intactness of the unedited regions.

For the human preference study, we collected questionnaire responses from 25 independent participants. Evaluators were presented with the original source image alongside the specific editing instruction, and were asked to rate the generated outputs from different models on a scale ranging from 0 to 5, where higher scores indicate superior generation quality. The final human preference score for each method is computed as the average rating across all participants and test

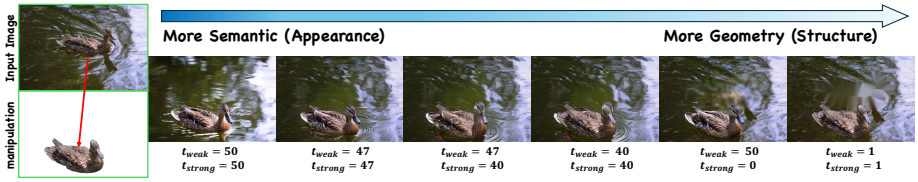

Fig. 6: Visual Ablation on Timestep Thresholds. We illustrate the fundamental trade-off between semantic realism (left) and geometric structure preservation (right). Relying predominantly on the generative prior (e.g., $t_w = 50, t_s = 50$) grants excessive freedom, resulting in structural deviation from the proxy. Conversely, excessive proxy injection at low noise levels (e.g., $t_w = 1, t_s = 1$) rigidly preserves geometry but introduces severe rendering artifacts and blending failures. Our optimal configuration ($t_w = 47, t_s = 40$) strikes the perfect balance, achieving seamless background harmonization while strictly adhering to the intended 3D spatial transformation.

Fig. 7: Qualitative Ablation on Proposed modules. We demonstrate the visual impact of each core module. The Naive Baseline struggles with both 3D skeleton preservation and background fidelity, yielding a distorted pose and altered context. Removing the warm-start initialization (w/o Warm-Start) results in an unnatural background synthesis, failing to smoothly harmonize the generated textures with the original scene. Critically, injecting uncalibrated spatial constraints (w/o Variance Homogeneity) severely disrupts the homogeneous latent distribution, leading to catastrophic rendering artifacts. In contrast, Ours seamlessly executes the complex spatial manipulation while preserving a highly natural and faithful background.

cases, achieving a Krippendorff's α of 0.74 which indicates substantial inter-rater agreement.

To complement the human study and provide a scalable complementary assessment, we employ a Gemini-family multimodal model [57] as an expert VLM judge. Prompted to act as an impartial image quality scorer, the VLM is provided with the identical input trio—the source image, the manipulation instruction, and the generated result. It analyzes visual coherence and instruction alignment to assign a comprehensive quality score from 0 to 5. As reported in Tab. 1, our framework secures the highest preference scores from both human evaluators and the AI judge, corroborating its robust perceptual superiority (inter-rater agreement details in Suppl. Material).

4.4 Ablation Studies

Effectiveness of Dual-Branch Denoising. To analyze the effectiveness of our dual-branch denoising framework, we conduct ablation studies on the critical

timestep thresholds summarized in Tab. 2. The results reveal a clear trade-off between geometric structure preservation and semantic realism under conventional editing settings. As shown in Fig. 6, relying on single-timestep initialization struggles to balance foreground geometry and background generation. A high initialization noise at step 47 increases generative freedom and achieves a CLIP score of 0.921, but weakens geometric fidelity, leading to a lower PSNR of 20.724 and a DreamSim error of 0.044. Reducing the initialization noise to step 40 improves spatial alignment (PSNR 21.822) but limits semantic diversity, lowering the CLIP score to 0.906.

Extreme configurations further highlight this limitation. Initializing from pure noise destroys the prescribed geometry (PSNR 18.479), while excessive proxy injection exposes rendering artifacts and degrades feature consistency (DINO 0.812, DreamSim 0.084). Moreover, enforcing full injection across all timesteps disrupts the diffusion process, reducing the DINO score to 0.860.

In contrast, our calibrated denoising window ($t_w = 47$, $t_s = 40$) achieves the best balance between structural fidelity and semantic realism, improving PSNR to 23.499 and DINO to 0.961 while reducing DreamSim to 0.027, validating the effectiveness of our asymmetric constraint strategy.

Effectiveness of Variance-Homogeneous Injection. Variance-homogeneous injection serves as the mathematical cornerstone of our spatial constraint mechanism. As reported in Tab. 3, replacing our synchronized forward-noise injection with uncalibrated constraints (w/o Variance Homogeneity) leads to a catastrophic performance collapse. Disrupting the homogeneous latent distribution of the pre-trained diffusion model severely degrades deep feature representations, causing the DINO score to plummet from 0.961 to 0.314 and the CLIP score to drop to 0.682. Structurally, this lack of variance alignment induces severe geometric distortions, increasing the DreamSim error to 0.240. This quantitative degradation is vividly reflected in Fig. 7, where the absence of this module produces extreme attention leakage and catastrophic rendering artifacts, highlighting that variance homogeneity is critical for stable generative editing.

Table 3: Quantitative Ablation on Proposed Components. We validate our framework using a leave-one-out strategy on the unified 50-pair protocol. Removing any core component significantly degrades either geometric preservation or semantic harmony, while our full model achieves the optimal balance across all metrics.

Model	PSNR ↑	DINO ↑	CLIP ↑	LPIPS ↓	DreamSim ↓
Naive Baseline	16.217	0.835	0.890	0.253	0.049
w/o Warm-Start	21.320	0.943	0.942	0.134	0.043
w/o Variance Homogeneity	11.494	0.314	0.682	0.520	0.240
Ours (Full Model)	**23.499**	**0.961**	**0.952**	**0.114**	**0.027**

Effectiveness of Warm-Start Initialization. The warm-start initialization is crucial for anchoring the global color layout and preserving the unedited background context. When this module is removed (w/o Warm-Start), the diffusion process is forced to reconstruct the scene from pure Gaussian noise without the low-frequency guidance of the source image. Tab. 3 demonstrates that this omission causes a decline in pixel-level and perceptual fidelity, with PSNR dropping

from 23.499 to 21.320 and the LPIPS perceptual error surging from 0.114 to 0.134. Qualitatively, as demonstrated in Fig. 7, failing to warm-start the latents results in an unnatural background synthesis that completely alters the original illumination and environmental context. By successfully integrating this initialization, our full model reliably executes complex spatial manipulations while maintaining a highly faithful background.

5 Conclusion

In this paper, we presented **GeoEdit**, a principled framework for enforcing strict 3D physical constraints, including translation, rotation, and scaling, in single-image object manipulation. We show that the core challenge in diffusion-based editing arises from the asymmetric generative requirements of the scene: geometry-constrained foreground manipulation and free-form background hallucination. To address this, GeoEdit adopts a *Lift-Manipulate-Render-Denoise* pipeline that lifts 2D editing into a geometry-aware 3D space. Our Dual-Branch Denoising mechanism bridges the rendered proxy and diffusion prior through variance-consistent injection, effectively preventing attention leakage and ghosting artifacts. We further introduce **GeoEditBench**, a comprehensive benchmark for evaluating spatial transformations. Extensive experiments demonstrate that GeoEdit achieves state-of-the-art performance in geometric accuracy, identity preservation, and background harmonization.

6 Limitations

While GeoEdit demonstrates precise 3D-aware editing, its modular pipeline introduces several limitations. First, **error propagation**: inaccuracies in depth estimation or monocular 3D lifting can distort the reconstructed geometry, leading to structural errors in the rendered proxy that propagate through subsequent stages. Second, **extreme spatial manipulations** remain challenging. Large object translations may expose disoccluded regions beyond the capacity of the background prior, causing inpainting failures, while large rotations often require hallucinating previously unseen back-facing surfaces, which can result in blurred or inconsistent textures. Third, **computational cost**: we employ the native 81-frame context of the video backbone to fully exploit its temporal prior for rigid 3D consistency; reducing this context degrades structural integrity, as shown in the Supplementary Material. Although this improves structural integrity, it increases inference time compared with purely 2D editing approaches; future work will explore lightweight fine-tuning and more efficient architectures to reduce this requirement. Finally, accurately reproducing complex view-dependent lighting, shadows, and specular effects remains difficult because these phenomena are synthesized primarily through learned generative priors rather than explicit physical modeling.

7 Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62276154), the Natural Science Foundation of Guangdong Province (Grant No. 2024TQ08X729), the Basic Research Fund of Shenzhen City (Grant Nos. JCYJ20240813112009013 and GJHZ20240218113603006), and the Major Key Project of Peng Cheng Laboratory for Experiments and Applications (Grant No. PCL2024A08).

References

1. Gemini 2.5 flash image (nano banana) | google ai studio. https://aistudio.google.com/models/gemini-2-5-flash-image
2. 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. pp. 18208–18218 (2022)
3. Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv e-prints pp. arXiv–2506 (2025)
4. 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)
5. Cao, W., Zhang, H., Tian, F., Wu, Y., Li, Y., Wang, S., Yu, N., Liu, Y.: Freeorbit4d: Training-free arbitrary camera redirection for monocular videos via geometry-complete 4d reconstruction. arXiv preprint arXiv:2601.18993 (2026)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
7. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
8. Chen, J., Mehran, R., Jia, X., Xie, S., Woo, S.: Blenderfusion: 3d-grounded visual editing and generative compositing. arXiv preprint arXiv:2506.17450 (2025)
9. Chen, Y., He, X., Ma, X., Ma, Y.: Contextflow: Training-free video object editing via adaptive context enrichment. arXiv preprint arXiv:2509.17818 (2025)
10. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. arXiv preprint arXiv:2210.11427 (2022)
11. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13142–13153 (2023)
12. Feng, K., Ma, Y., Wang, B., Qi, C., Chen, H., Chen, Q., Wang, Z.: Dit4edit: Diffusion transformer for image editing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2969–2977 (2025)
13. Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: Dreamsim: Learning new dimensions of human visual similarity using synthetic data. arXiv preprint arXiv:2306.09344 (2023)

14. Gao, H., Tang, B., Song, Y., Fang, G., He, Z., Yang, J., Shou, M.Z.: Pai-studio: Cinematic video background replacement with camera-aware motion. arXiv preprint arXiv:2606.01399 (2026)
15. He, X., Liu, Q., Qian, S., Wang, X., Hu, T., Cao, K., Yan, K., Zhang, J.: Id-animator: Zero-shot identity-preserving human video generation. arXiv preprint arXiv:2404.15275 (2024)
16. He, X., Liu, Q., Ye, Z., Ye, W., Wang, Q., Wang, X., Chen, Q., Wan, P., Zhang, D., Gai, K.: Fulldit2: Efficient in-context conditioning for video diffusion transformers. arXiv preprint arXiv:2506.04213 (2025)
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
18. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
19. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17191–17202 (2025)
20. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
21. Li, R., Zhang, H., Zhang, Y., Zhang, Y., Zhang, Y., Guo, J., Zhang, Y., Li, X., Liu, Y.: Lodge++: High-quality and long dance generation with robust choreography patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
22. Li, R., Zhang, Y., Zhang, Y., Zhang, H., Guo, J., Zhang, Y., Liu, Y., Li, X.: Lodge: A coarse to fine diffusion network for long dance generation guided by the characteristic dance primitives. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1524–1534 (2024)
23. Li, R., Zhao, J., Zhang, Y., Su, M., Ren, Z., Zhang, H., Tang, Y., Li, X.: Finedance: A fine-grained choreography dataset for 3d full body dance generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10234–10243 (2023)
24. Liu, R., Wu, R., Hoorick, B.V., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object (2023), https://arxiv.org/abs/2303.11328
25. Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., Wang, W.: Syncdreamer: Generating multiview-consistent images from a single-view image. arXiv preprint arXiv:2309.03453 (2023)
26. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11461–11471 (2022)
27. Ma, Y., Cun, X., Liang, S., Xing, J., He, Y., Qi, C., Chen, S., Chen, Q.: Magicstick: Controllable video editing via control handle transformations. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 9385–9395. IEEE (2025)
28. Ma, Y., Feng, K., Hu, Z., Wang, X., Wang, Y., Zheng, M., He, X., Zhu, C., Liu, H., He, Y., et al.: Controllable video generation: A survey. arXiv preprint arXiv:2507.16869 (2025)
29. Ma, Y., Feng, K., Zhang, X., Liu, H., Zhang, D.J., Xing, J., Zhang, Y., Yang, A., Wang, Z., Chen, Q.: Follow-your-creation: Empowering 4d creation through video inpainting. arXiv preprint arXiv:2506.04590 (2025)

30. Ma, Y., He, Y., Cun, X., Wang, X., Chen, S., Li, X., Chen, Q.: Follow your pose: Pose-guided text-to-video generation using pose-free videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4117–4125 (2024)
31. Ma, Y., He, Y., Wang, H., Wang, A., Shen, L., Qi, C., Ying, J., Cai, C., Li, Z., Shum, H.Y., et al.: Follow-your-click: Open-domain regional image animation via motion prompts. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 6018–6026 (2025)
32. Ma, Y., Liu, H., Wang, H., Pan, H., He, Y., Yuan, J., Zeng, A., Cai, C., Shum, H.Y., Liu, W., et al.: Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–12 (2024)
33. Ma, Y., Liu, Y., Zhu, Q., Yang, A., Feng, K., Zhang, X., Li, Z., Han, S., Qi, C., Chen, Q.: Follow-your-motion: Video motion transfer via efficient spatial-temporal decoupled finetuning. arXiv preprint arXiv:2506.05207 (2025)
34. Ma, Y., Wang, X., Ma, Q., Wang, Q., Zheng, M., Yang, X., Li, H., Zhao, C., Ying, J., Yang, H., et al.: Group editing: Edit multiple images in one go. arXiv preprint arXiv:2603.22883 (2026)
35. Ma, Y., Wang, Z., Ren, T., Zheng, M., Liu, H., Guo, J., Fong, M., Xue, Y., Zhao, Z., Schindler, K., et al.: Fastvmt: Eliminating redundancy in video motion transfer. arXiv preprint arXiv:2602.05551 (2026)
36. Ma, Y., Yan, Z., Liu, H., Wang, H., Pan, H., He, Y., Yuan, J., Zeng, A., Cai, C., Shum, H.Y., et al.: Follow-your-emoji-faster: Towards efficient, fine-controllable, and expressive freestyle portrait animation. arXiv preprint arXiv:2509.16630 (2025)
37. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
38. Michel, O., Bhattad, A., VanderBilt, E., Krishna, R., Kembhavi, A., Gupta, T.: Object 3dit: Language-guided 3d-aware image editing. Advances in Neural Information Processing Systems **36**, 3497–3516 (2023)
39. Nan, K., Zhao, W., Zhou, P., Li, J., Yang, Z., Yang, J., Tai, Y.: Accelerating autoregressive video diffusion via history-guided cache and residual correction. In: CVPR. pp. 43740–43750 (2026)
40. Pan, X., Tewari, A., Leimkühler, T., Liu, L., Meka, A., Theobalt, C.: Drag your gan: Interactive point-based manipulation on the generative image manifold. In: ACM SIGGRAPH 2023 conference proceedings. pp. 1–11 (2023)
41. Pandey, K., Guerrero, P., Gadelha, M., Hold-Geoffroy, Y., Singh, K., Mitra, N.J.: 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. pp. 7695–7704 (2024)
42. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2085–2094 (2021)
43. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023)
44. Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: Fatezero: Fusing attentions for zero-shot text-based video editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15932–15942 (2023)
45. 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. PmLR (2021)

46. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021)
47. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
48. Ruan, P., Zi, B., Qi, X., Huang, Y., Xiao, R., Wang, P., Cao, J., Shi, Y.: Ctrl&shift: High-quality geometry-aware object manipulation in visual generation. arXiv preprint arXiv:2602.11440 (2026)
49. Sajnani, R., Vanbaar, J., Min, J., Katyal, K.D., Sridhar, S.: Geodiffuser: Geometry-based image editing with diffusion models. In: Proceedings of the Winter Conference on Applications of Computer Vision. pp. 472–482 (2025)
50. Shen, Y., Gu, J., Tang, X., Zhou, B.: Interpreting the latent space of gans for semantic face editing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9243–9252 (2020)
51. Shen, Y., Yuan, J., Aonishi, T., Nakayama, H., Ma, Y.: Follow-your-preference: Towards preference-aligned image inpainting. arXiv preprint arXiv:2509.23082 (2025)
52. Shi, Y., Xue, C., Liew, J.H., Pan, J., Yan, H., Zhang, W., Tan, V.Y., Bai, S.: Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8839–8849 (2024)
53. Singer, A., Rotstein, N., Mann, A., Kimmel, R., Litany, O.: Time-to-move: Training-free motion controlled video generation via dual-clock denoising. arXiv preprint arXiv:2511.08633 (2025)
54. Song, Y., Huang, S., Yao, C., Ci, H., Ye, X., Liu, J., Zhang, Y., Shou, M.Z.: Processpainter: Learning to draw from sequence data. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–10 (2024)
55. Song, Y., Liu, C., Jiang, Y., Shou, M.Z.: Streamingeffect: Real-time human-centric video effect generation. arXiv preprint arXiv:2605.17019 (2026)
56. Song, Y., Yao, W., Wang, H., Shou, M.Z.: Vista: Triplet-supervised video style transfer with diffusion transformers. arXiv preprint arXiv:2605.17312 (2026)
57. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
58. Telea, A.: An image inpainting technique based on the fast marching method. Journal of graphics tools **9**(1), 23–34 (2004)
59. Voleti, V., Yao, C.H., Boss, M., Letts, A., Pankratz, D., Tochilkin, D., Laforte, C., Rombach, R., Jampani, V.: Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In: European Conference on Computer Vision. pp. 439–457. Springer (2024)
60. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
61. Wang, J., Ma, Y., Guo, J., Xiao, Y., Huang, G., Li, X.: Cove: Unleashing the diffusion feature correspondence for consistent video editing. Advances in Neural Information Processing Systems **37**, 96541–96565 (2024)
62. 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)

63. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
64. Wiles, O., Gkioxari, G., Szeliski, R., Johnson, J.: Synsin: End-to-end view synthesis from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7467–7477 (2020)
65. 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)
66. Xie, S., Tu, Z.: Holistically-nested edge detection (2015), https://arxiv.org/abs/1504.06375
67. Xu, X., Li, Y., You, S., Bao, B.K.: Smrabooth: Subject and motion representation alignment for customized video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16130–16141 (2026)
68. Yang, H., Hong, L., Li, A., Hu, T., Li, Z., Lee, G.H., Wang, L.: Contranerf: Generalizable neural radiance fields for synthetic-to-real novel view synthesis via contrastive learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16508–16517 (2023)
69. Yang, X., Xie, J., Yang, Y., Huang, Y., Xu, M., Wu, Q.: Unified video editing with temporal reasoner. arXiv preprint arXiv:2512.07469 (2025)
70. Ye, Z., He, X., Liu, Q., Wang, Q., Wang, X., Wan, P., Zhang, D., Gai, K., Chen, Q., Luo, W.: Unic: Unified in-context video editing. ICLR 2026 (2025)
71. Yenphraphai, J., Pan, X., Liu, S., Panozzo, D., Xie, S.: Image sculpting: Precise object editing with 3d geometry control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4241–4251 (2024)
72. Yu, X., Wang, T., Kim, S.Y., Guerrero, P., Chen, X., Liu, Q., Lin, Z., Qi, X.: Objectmover: Generative object movement with video prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17682–17691 (2025)
73. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
74. Zhang, Q., Xu, Y., Wang, C., Lee, H.Y., Wetzstein, G., Zhou, B., Yang, C.: 3ditscene: Editing any scene via language-guided disentangled gaussian splatting. In: International Conference on Learning Representations. vol. 2025, pp. 2760–2775 (2025)
75. Zhang, Z., Lin, M., Song, Q., Zhang, Y., Ji, R.: Objectadd: adding objects into image via a training-free diffusion modification fashion. Pattern Recognition p. 112807 (2025)
76. Zhao, W., Han, Y., Tang, J., Wang, K., Luo, H., Song, Y., Huang, G., Wang, F., You, Y.: Dydit++: Diffusion transformers with timestep and spatial dynamics for efficient visual generation. TPAMI (2026)
77. Zhao, W., Han, Y., Tang, Z., Tang, J., et al.: Rapidˆ 3: Tri-level reinforced acceleration policies for diffusion transformer. In: ICLR (2026)