UniGeo: Unifying Geometric Guidance for Camera-Controllable View Synthesis via Video Models

Hong Jiang[1], Wensong Song[1], Zongxin Yang[2], Ruijie Quan[1], and Yi Yang[1⋆]

[1] ReLER, CCAI, College of Artificial Intelligence, Zhejiang University, China
[2] DBMI, HMS, Harvard University, USA
Project Page: https://mo230761.github.io/UniGeo.github.io/

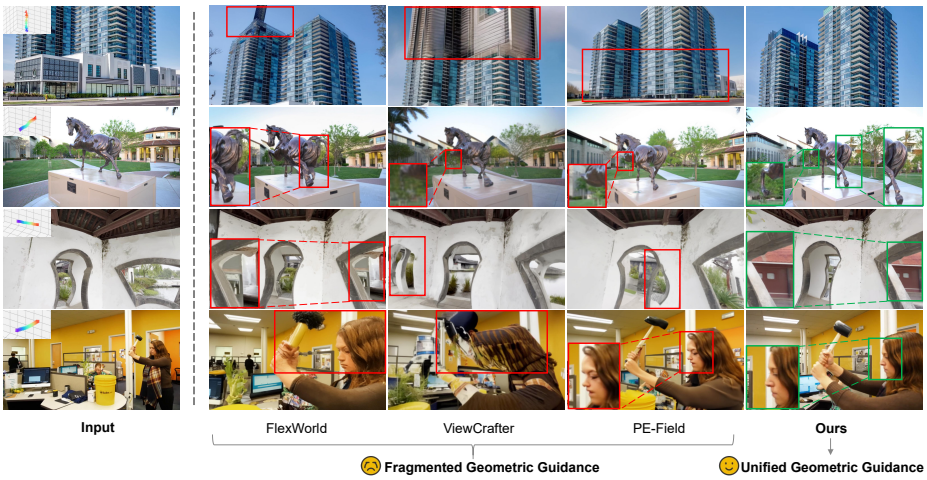

Fig. 1: Visual comparisons. Existing methods relying on fragmented geometric guidance often suffer from structural distortions or artifacts under camera motion (highlighted in red). In contrast, by enforcing unified geometric guidance, our **UniGeo** successfully preserves global scene geometry and structural fidelity (highlighted in green, with selected details enlarged).

Abstract. Camera-controllable view synthesis aims to synthesize novel views of a given scene under varying camera poses while strictly preserving cross-view geometric consistency. However, existing methods typically rely on fragmented geometric guidance, such as only injecting point clouds at the representation level despite models containing multiple levels, and are mainly based on image diffusion models that operate on discrete view mappings. These two limitations jointly lead to geometric drift and structural degradation under continuous camera motion. We observe that while leveraging video models provides continuous viewpoint priors for camera-controllable view synthesis, they still struggle to form stable

⋆ Corresponding author.

geometric understanding if geometric guidance remains fragmented. To systematically address this, we inject unified geometric guidance across the three levels that jointly determine the generative output: representation, architecture, and loss function. To this end, we propose **UniGeo**, a novel camera-controllable editing framework. Specifically, at the representation level, UniGeo incorporates a frame-decoupled geometric reference injection mechanism to provide robust cross-view geometry context. Furthermore, at the architecture level, it introduces a geometric anchor attention to align multi-view features, and at the loss function level, it proposes a trajectory-endpoint geometric supervision strategy to explicitly reinforce the structural fidelity of target views. Experiments across multiple public benchmarks, encompassing both extensive and limited camera motion settings, demonstrate that UniGeo significantly outperforms existing methods in visual quality and geometric consistency.

Keywords: Camera-controllable view synthesis · Unified geometric guidance · Video Models

1 Introduction

Camera-controllable view synthesis [7, 16, 50, 53, 61, 62] aims to generate transformations of the scene under different viewpoints given changes in camera pose, while maintaining strict cross-view geometric consistency. This capability is critical for applications such as film post-production and robotic perception, among others, directly affecting rendering quality and perception reliability. Unlike general image editing [3, 12, 20, 41, 54, 55, 70, 94], the core challenge here lies not simply in modifying appearance attributes, but in preserving scene structure across views with consistent and seamless generation.

High-quality camera-controllable view synthesis requires a generative framework that strictly follows camera motion rules while preserving geometric consistency. However, existing methods still face significant challenges: **(i) Lack of Continuity.** Camera motion is continuous in 3D space, and an ideal framework should reflect the continuous evolution of the scene along the camera trajectory. However, most existing methods [7, 50, 61, 62] are based on image diffusion models and only target mappings between discrete viewpoints, lacking the ability to model continuous camera trajectories. This often results in unstable generation. **(ii) Lack of Unified Geometric guidance.** Real-world viewpoint changes share consistent geometric correspondences, requiring the model to possess unified guidance over the generation process. However, existing methods [18, 28, 89] typically enforce fragmented geometric guidance (such as only injecting point clouds or depth at the representation level). This fragmentation confines geometric guidance to an isolated level, leaving the rest of the model disjointed and unable to form unified correspondences. Consequently, this results in breakages in geometric guidance propagation and ultimately leads to 3D structure collapse.

Based on these observations, we note that video models naturally possess continuous-viewpoint modeling capabilities, providing a potential foundation for

camera-controllable view synthesis [59,76]. However, even within video models, if geometric guidance is fragmented, the network struggles to form stable geometric understanding across different views [18, 30, 73, 89]. To systematically address this issue, we draw inspiration from the fundamental levels of model design: representation, architecture, and loss function [9, 23, 27, 44]. Since these three levels jointly determine the generation process, this naturally motivates our framework: to ensure global geometric consistency, we systematically inject geometric guidance into each of them.

Motivated by this analysis, we propose **UniGeo**. Unlike existing methods [7, 18, 26, 89] that enforce fragmented geometric guidance, UniGeo incorporates unified geometric guidance across representation, architecture, and loss function, systematically rethinking the use of video models for camera-controllable view synthesis. Specifically, UniGeo achieves this through three tightly coupled modules: **(i) Frame-Decoupled Point Cloud Injection.** Point clouds encode scene geometry and cross-view correspondences, serving as effective priors. At the representation level, we lift the input image into a trajectory-aligned point cloud sequence and inject it into a video model. Unlike prior methods [28, 89, 91] that concatenate point clouds along the channel dimension, we decouple them along the frame dimension into independent geometric reference frames. This avoids forcing a strict pixel-to-pixel alignment, preventing the inherently missing points in point clouds from directly corrupting the generated image. **(ii) Geometric Anchor Attention.** At the architecture level, we introduce an attention mechanism using first-frame geometric features as *geometry anchors*. Unlike existing I2V methods that only maintain appearance continuity [67, 84], our geometric anchors focus on preserving unified geometric consistency across views. During attention interactions, the anchors continuously align features from different viewpoints. **(iii) Trajectory-Endpoint Geometric Supervision.** At the loss function level, we propose a geometric supervision strategy focusing on camera trajectory endpoints. This shifts the optimization objective from sequence-level consistency to structural fidelity in the target viewpoint. With sparse temporal sampling of key viewpoints, this strategy reduces over-modeling of intermediate frames and applies higher geometric weights to trajectory endpoints, strengthening constraints on target-view 3D structures.

We conduct comprehensive experiments on multiple public video datasets, including RealEstate10K (RE10K) [99], Tanks and Temples (Tanks) [38], DL3DV [48], among others. Unlike previous approaches [18, 58, 89] that split test sets according to video frame intervals, we partition the test videos based on the proportion of newly synthesized regions, categorizing them into *extensive* and *limited camera motion settings*. On RE10K videos with extensive camera motion, the LPIPS decreases from 0.3008 to 0.2377, and on Tanks videos with limited camera motion, the PSNR increases from 16.9580 to 17.8171. These results demonstrate that unified geometric guidance effectively improves cross-view geometric consistency and substantially strengthens camera-controllable view synthesis across diverse camera motions.

In summary, the main contributions of this work are as follows:

- To the best of our knowledge, we propose UniGeo, the first camera-controllable editing framework centered on unified geometric guidance. By overcoming the limitations of relying solely on fragmented geometric guidance and leveraging the continuous-viewpoint prior of video models, our method achieves state-of-the-art (SOTA) performance in camera-controllable view synthesis under both extensive and limited camera motion settings.
- To instantiate the unified geometric guidance, we systematically design a tightly coupled model. This comprises, at the representation level, a **frame-decoupled point cloud injection** mechanism to provide robust geometric context; at the architecture level, a **geometric anchor attention** module to ensure cross-view structural consistency; and at the loss function level, a **trajectory-endpoint geometric supervision** strategy to explicitly strengthen the 3D structural fidelity of target viewpoints.

2 Related Work

Image Editing. Image editing aims to modify visual content in a controllable manner [34]. Early diffusion-based approaches rely on training-free inversion [12,20,31,36,54–56,60,63,82] and model fine-tuning [8,13,35,37,88,90,95]. Recently, this field has been further advanced by large-scale text-to-image foundation models [3, 25, 41, 42, 64, 70, 86, 94, 100] and unified auto-regressive architectures [17, 21, 22, 46, 47, 51, 69, 74, 75, 79] for fine-grained semantic control. While excelling at appearance manipulation, these approaches generally lack explicit spatial viewpoint control. Moving beyond general semantic editing, recent works explore camera-controllable view synthesis [7, 16, 26, 50, 53, 61, 62, 78, 81]. However, as these methods use image diffusion models, they struggle to stably model continuous camera motion, causing geometric inconsistencies across views.

Video Priors for Image Editing. With the rapid progress of video generation models [2, 10, 14, 24, 39, 57, 68, 98], recent studies have explored directly adapting pretrained video models for image editing tasks. These efforts aim to leverage the continuous temporal priors of video models to maintain structural and semantic consistency during the editing process. Unlike conventional formulations that treat image editing as an independent single-frame generation problem, these approaches directly employ video models to perform image manipulation [59, 76, 93]. However, existing methods generally lack geometric guidance and fail to model the relationship between camera motion and 3D structures, which fundamentally limits their ability to support camera-controllable view synthesis.

Camera-Controllable Video Generation. Camera-controllable video generation [5,29,40,65,83,96,97] incorporates camera motion into video models, enabling synthesized videos to exhibit continuous and consistent viewpoint changes. Existing studies typically introduce conditional signals into the generation process, including encoded camera parameters [4, 6, 30, 66, 73, 80], monocular depth [15, 28], or view warping [18, 33, 58, 87, 89]. These signals can be applied to pretrained video models, providing temporal and spatial guidance. However, the geometric guidance in these methods is typically fragmented. They lack unified geometric

guidance across the generation pipeline, thus struggle to maintain structural fidelity under continuous camera motion.

3 Background: Rectified Flow for Video Diffusion Models

Modern video generation models typically employ a 3D-VAE to compress raw videos into a compact latent space, improving computational efficiency [11,77,84]. Given an input video x, the VAE encoder extracts its latent representation $z_0 = E(x)$, where all subsequent generative modeling is performed. Finally, a decoder reconstructs the generated latents $\hat{z}_0$ back to the pixel space as $\hat{x} = D(\hat{z}_0)$.

Unlike standard diffusion models, recent video foundation models (e.g., Wan [67]) learn the latent distribution via a rectified flow formulation based on flow matching [1,25,49,52]. For a latent data sample z_0 and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, an intermediate state z_t at continuous time $t \in [0, 1]$ is constructed by linearly blending the data latent and the noise as:

$$z_t = (1 - t)z_0 + t\epsilon, \tag{1}$$

A neural network F_θ is trained to predict the target velocity field $v = \epsilon - z_0$ that drives this flow. Taking the latent state z_t, the timestep t, and the conditioning signals (text y and image c) as inputs, the training objective is simplified to:

$$\mathcal{L}_\theta = \mathbb{E}_{t,x,\epsilon} \left\| F_\theta([z_t, y, c], t) - (\epsilon - z_0) \right\|_2^2 . \tag{2}$$

4 UniGeo Model

Given an input image $I_0 \in \mathbb{R}^{3\times H\times W}$ and a camera control prompt, our goal is to synthesize a novel view under the target camera pose, while preserving the 3D geometric structure. To this end, we systematically introduce unified geometric guidance at multiple levels to enable accurate camera-controllable view synthesis.

As shown in Fig. 2, our approach consists of three modules: Section 4.1 introduces **Frame-Decoupled Point Cloud Injection**, which constructs point cloud features and injects them into the video model along the frame dimension, including two components: *Point Cloud Geometry Construction* and *Frame-Decoupled Geometry Injection*; Section 4.2 presents **Geometric Anchor Attention**, which continuously aligns geometric features across views; and Section 4.3 describes **Trajectory-Endpoint Geometric Supervision**, guiding the model to maintain geometric structures at target viewpoints.

4.1 Frame-Decoupled Point Cloud Injection

Point Cloud Geometry Construction. Directly injecting camera parameters as geometric guidance into diffusion models [6, 30, 73] forces the network to implicitly learn the mapping from camera poses to appearance changes. This strategy usually provides only coarse camera controllability and exhibits limited

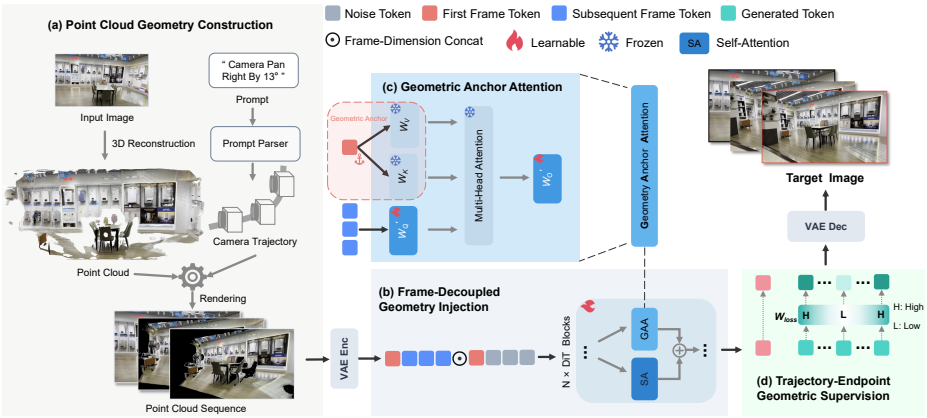

Fig. 2: UniGeo Framework. UniGeo incorporates unified geometric guidance through: (a) Geometry Construction: Lifting input images into 3D point cloud sequences. (b) Frame-Decoupled Geometry Injection: Injecting sequences along the frame dimension. (c) Geometric Anchor Attention: Aligning cross-view features using first-frame tokens as anchors. (d) Trajectory-Endpoint Geometric Supervision: Applying higher loss weights to trajectory endpoints versus intermediate frames.

generalization to unseen camera trajectories. Inspired by prior work [33, 87, 89], we introduce point cloud sequences as geometric guidance, supplying the video generation model with explicit geometric priors (as shown in Fig. 2(a)) .

During training, given an input video $V = [I_0, \ldots, I_{N-1}] \in \mathbb{R}^{N\times 3\times H\times W}$, where N denotes the number of frames, we first employ VGGT [71] to estimate the camera pose of each frame, yielding the camera trajectory $\mathcal{C} = \{C_0, \ldots, C_{N-1}\}$. Meanwhile, a point cloud P_0 is reconstructed from the first frame I_0 using the pre-trained VGGT model [71].

We then move the virtual camera along the estimated camera trajectory $\mathcal{C} = \{C_0, \ldots, C_{N-1}\}$ and render the point cloud to obtain a sequence of renderings:

$$R_f = \pi(P_0, C_f), \quad f = 0, \ldots, N-1, \tag{3}$$

where C_f denotes the camera pose at the f-th timestep along the target camera trajectory, and $\pi(\cdot)$ is a differentiable rendering operator. To ensure accurate alignment at the reference view, we explicitly replace the first rendered frame with the original high-fidelity input image, i.e., $R_0 = I_0$.

This procedure produces a rendering sequence that is aligned with the target camera trajectory and serves as geometric guidance for subsequent video generation. Notably, since both the point cloud and the camera poses are estimated by the same model, they naturally reside in a unified coordinate and scale space, which eliminates potential scale inconsistency issues.

During the inference process, we first convert the user-specified camera control instructions, which indicate the desired camera motion, into the corresponding target camera pose for the ending view. Then, we interpolate the camera trajectory

at uniform intervals between the initial camera pose and the target ending pose to obtain a sequence of camera poses, which is subsequently used to render the point cloud and produce the corresponding rendering sequence.

Frame-Decoupled Geometry Injection. Motivated by the context conditioning design in DiT-based models [6, 64, 94], we introduce the rendered point cloud sequence as *frame-decoupled* geometric context and inject it into the video diffusion model by concatenating it with target video tokens (as shown in Fig. 2(b)).

Specifically, let z_t denote the latent representation of the target video and z_s denote the latent representation of the rendered point cloud sequence. We apply a patchification operation to obtain $x_t = \text{patchify}(z_t)$ and $x_s = \text{patchify}(z_s)$. These tokens are then concatenated along the frame dimension:

$$x_i = [x_t, x_s]_{\text{frame-dim}} \in \mathbb{R}^{b\times 2f\times s\times d}, \tag{4}$$

which is fed into the DiT backbone as the input token sequence.

This frame-decoupled injection design mitigates the adverse effects of imperfect point cloud priors and allows the geometric context to interact flexibly with target video features throughout the network, naturally supporting our unified guidance and thereby improving cross-view geometric consistency.

4.2 Geometric Anchor Attention

To maintain cross-view geometric consistency, we introduce **Geometric Anchor Attention**, which aligns features across different timesteps using the structural features of the first frame (as shown in Fig. 2(c)).

Specifically, given a video sequence of length N, we designate the first frame as the geometric anchor and let X_0 denote its intermediate feature representation extracted from the backbone network. Its corresponding key K_0 and value V_0 are obtained via pre-trained projection matrices W_K and W_V, i.e., $K_0 = X_0 W_K$ and $V_0 = X_0 W_V$. For any subsequent frame $i \in \{1, \ldots, N-1\}$ with features X_i, we derive a specific geometric query $(Q_i)' = X_i W'_Q$, where W'_Q is a trainable weight matrix initialized directly from the pre-trained values of the corresponding layer. The Geometric Anchor Attention is then defined as:

$$\text{Attention}((Q_i)', K_0, V_0) = \text{softmax}\Big(\frac{(Q_i)' K_0^\top}{\sqrt{d}}\Big) V_0. \tag{5}$$

The final feature representation is obtained by summing the original self-attention output and the proposed Geometric Anchor Attention:

$$X_i^{\text{out}} = \text{Attention}(Q_i, K_i, V_i) W_O + \alpha \cdot \text{Attention}((Q_i)', K_0, V_0) W'_O, \tag{6}$$

where Q_i, K_i, V_i denote the queries, keys, and values of the original attention mechanism, and W_O is its pre-trained output projection. To ensure training stability and preserve the original generative prior, the new projection matrix W'_O is zero-initialized. Additionally, a scalar weight α is introduced to explicitly control the influence of the geometric guidance.

Table 1: Quantitative comparison of our model with relevant methods under the **extensive camera motion** setting demonstrates that our model substantially surpasses all relevant baselines across all key metrics. The best and second-best results are demonstrated in **bold** and underlined, respectively.

Method	DL3DV				RE10K				Tanks			
	FID↓	SSIM↑	LPIPS↓	PSNR↑	FID↓	SSIM↑	LPIPS↓	PSNR↑	FID↓	SSIM↑	LPIPS↓	PSNR↑
CameraCtrl [30]	172.63	0.2703	0.4822	10.0938	124.87	0.4320	0.4266	11.1563	116.02	0.3787	0.3922	11.4924
MotionCtrl [73]	151.00	0.3788	0.5242	10.2664	114.65	0.5228	0.4988	10.4614	103.97	0.4250	0.5077	10.0663
ViewCrafter [89]	146.81	0.4700	0.4556	12.6128	97.57	0.5905	0.3668	14.3176	105.99	0.5418	0.4047	13.3314
FlexWorld [18]	125.27	0.4766	0.3726	13.3029	90.43	0.6430	0.3008	14.3408	95.40	0.5476	0.3395	13.8118
PE-Field [7]	131.93	0.4534	0.4329	12.7060	105.34	0.6210	0.3768	13.1684	91.96	0.5222	0.3957	13.1063
Ours	**113.11**	**0.4830**	**0.3248**	**13.6067**	**66.67**	**0.6522**	**0.2377**	**14.9723**	**76.97**	**0.5574**	**0.2633**	**14.4537**

This design aligns features across timesteps using only two trainable matrices, W'_Q and W'_O. It adds minimal computational overhead while serving as the crucial feature-level component of our unified geometric guidance, thereby fundamentally improving cross-view structural consistency.

4.3 Trajectory-Endpoint Geometric Supervision

During temporal modeling, we apply **sparse temporal sampling** to uniformly select key frames, reducing computation on intermediate frames, and introduce **Trajectory-Endpoint Geometric Supervision**, which increases loss weights on trajectory-endpoint frames while reducing weights on intermediate frames to enforce geometric stability at the trajectory endpoints (as shown in Fig. 2(d)).

Formally, given a sequence of length N, we assign a temporally-varying loss weight to each subsequent frame i. To explicitly emphasize the trajectory endpoints, the weighting coefficient $w_{\text{loss}}(i)$ is defined as a quadratic function of the frame's normalized distance to the temporal center:

$$w_{\text{loss}}(i) = 1 + \gamma \left(\frac{2i}{N-1} - 1 \right)^2, \quad i = 1, \ldots, N-1, \tag{7}$$

where γ is a hyperparameter that controls the strength of this endpoint penalty.

The final weighted loss for the video sequence is then computed as:

$$\mathcal{L}_{\text{weighted}} = \sum_{i=1}^{N-1} w_{\text{loss}}(i) \mathcal{L}_i, \tag{8}$$

where $\mathcal{L}_i$ denotes the original flow matching loss at frame i.

To further strengthen geometric constraints at the target view, we adopt a **temporal extension** strategy at the end of the sequence, where the target-view frame is duplicated and extended to multiple consecutive timesteps for joint modeling. This design enforces persistent geometric guidance during the final stage of generation, ensuring a stable geometric structure at the target viewpoint.

Table 2: Quantitative comparison of our model with relevant methods under the **limited camera motion** setting demonstrates that our model substantially surpasses all relevant baselines across all key metrics. The best and second-best results are demonstrated in **bold** and underlined, respectively.

Method	DL3DV				RE10K				Tanks			
	FID↓	SSIM↑	LPIPS↓	PSNR↑	FID↓	SSIM↑	LPIPS↓	PSNR↑	FID↓	SSIM↑	LPIPS↓	PSNR↑
CameraCtrl [30]	141.83	0.2403	0.4252	10.5987	122.46	0.3994	0.3825	12.0042	97.27	0.3697	0.3266	13.0271
MotionCtrl [73]	124.60	0.3600	0.4835	10.7530	111.43	0.5110	0.4574	11.6774	84.60	0.4575	0.4257	12.0949
ViewCrafter [89]	103.26	0.5205	0.3619	15.2119	84.45	0.6370	0.2984	15.5421	73.97	0.5930	0.3124	16.1263
FlexWorld [18]	76.91	0.5049	0.2827	15.4140	73.80	0.6354	0.2573	16.1159	54.35	0.6016	0.2418	16.9580
PE-Field [7]	91.83	0.4662	0.3508	14.2761	81.76	0.6188	0.3189	15.4383	65.49	0.5681	0.3003	16.0821
Ours	**69.05**	**0.5271**	**0.2065**	**16.3740**	**51.73**	**0.6650**	**0.1730**	**17.2989**	**40.55**	**0.6278**	**0.1526**	**17.8171**

5 Experiments

5.1 Experimental Settings

Implementation Details. We adopt Wan2.2-TI2V-5B [67] as our base video generative model, fine-tuned with a LoRA of rank 256. During training, the frame resolution is fixed at 704×1248, and the video length is set to 29 frames, with the final four frames allocated for persistent modeling of the trajectory endpoints. The model is trained for approximately 10,000 iterations on 4 GPUs, with a learning rate of 1×10^{-4} and a total batch size of 4. The hyperparameters α for Geometric Anchor Attention and γ for Trajectory-Endpoint Geometric Supervision are set to 1 and 0.01, respectively.

Training Dataset. To ensure broad scene diversity, we utilize three large-scale datasets for training: DL3DV [48], MannequinChallenge [45], and RealEstate10K (RE10K) [99]. We curated approximately 3,500 samples from DL3DV, 2,500 from MannequinChallenge, and 9,000 from RE10K. Each selected sample consists of 81 frames, from which 29 frames are sparse-temporally sampled for training. All camera trajectories are consistently estimated using the pre-trained VGGT [71].

Testing and Evaluation. We evaluate our method on the test sets of RE10K, Tanks and Temples (Tanks) [38], DL3DV, and MannequinChallenge, which are also commonly used for dynamic scene understanding [85]. For RE10K, Tanks, and DL3DV, we categorize camera motion based on the proportion of newly synthesized regions (mask area) in the final frame of the point cloud rendering: videos with a mask ratio $> 35\%$ are classified as *extensive camera motion*, while the remainder are deemed *limited camera motion*. This 35% threshold is empirically chosen to distinguish significant viewpoint changes. We randomly select 50 video samples from each category. For MannequinChallenge, we randomly select 50 samples due to the inherent complexity of human-centric scenes. We adopt PSNR, SSIM [72], LPIPS [92], and FID [32] as primary evaluation metrics.

5.2 Comparisons with relevant methods

Quantitative comparisons. We evaluate our method against CameraCtrl [30], MotionCtrl [73], ViewCrafter [89], FlexWorld [18], and PE-Field [7] on DL3DV,

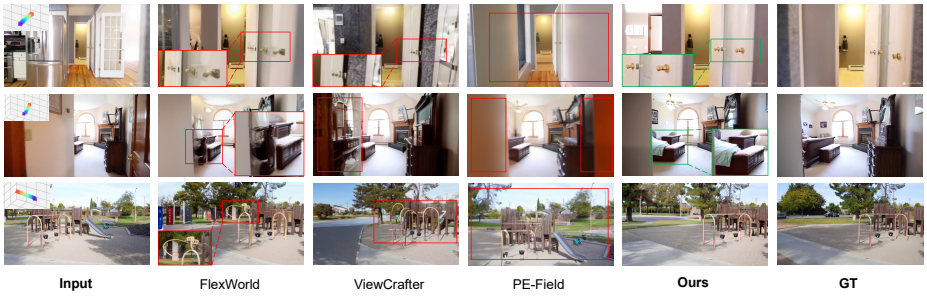

Fig. 3: Qualitative comparison under the **extensive camera motion** setting. Our approach better preserves the geometric structure of the scene.

Table 3: Quantitative comparison of temporal and geometric consistency on the **DL3DV dataset**. Our method achieves the best results under both extensive and limited camera motion, demonstrating superior cross-view geometric accuracy and temporal stability. The best and second-best results are **bold** and underlined, respectively.

Setting	Extensive Camera Variation			Limited Camera Variation		
	WarpErr ↓	RotErr ↓	TransErr ↓	WarpErr ↓	RotErr ↓	TransErr ↓
CameraCtrl [30]	0.3290	0.0592	0.2154	0.2800	0.0280	0.1609
MotionCtrl [73]	0.3094	0.2477	0.6227	0.2837	0.1180	0.6211
ViewCrafter [89]	0.2767	0.0320	0.0508	0.2545	<u>0.0118</u>	<u>0.0402</u>
FlexWorld [18]	<u>0.2708</u>	<u>0.0181</u>	<u>0.0442</u>	<u>0.2444</u>	0.0127	0.0505
Ours	**0.2666**	**0.0141**	**0.0310**	**0.2346**	**0.0075**	**0.0239**

RE10K, and Tanks under both extensive and limited camera motion settings, as shown in Tab. 1 and Tab. 2. Our method achieves the best performance across all key metrics, demonstrating strong capability to maintain high fidelity and perceptual quality under varying levels of camera motion. In addition, on the MannequinChallenge dataset [45] (Tab. 4), our method also attains the best results. To verify geometric consistency along continuous camera trajectories, we additionally report a temporal consistency metric (WarpErr [43]) and geometry-oriented metrics (RotErr / TransErr [6]) that explicitly measure cross-view geometric coherence. As shown in Tab. 3, our method yields the lowest errors across all these metrics, confirming its strong geometric and temporal consistency under camera motion.

Qualitative comparisons. Fig. 3 and Fig. 4 present qualitative comparisons under the extensive and limited camera motion settings, respectively. It can be observed that, under camera motion, existing methods often struggle to preserve the geometric structure of the scene during novel view generation, leading to artifacts such as

Table 4: Results on **MannequinChallenge**.

Method	FID↓	SSIM↑	LPIPS↓	PSNR↑
CameraCtrl [30]	207.43	0.3031	0.5316	9.7386
MotionCtrl [73]	200.98	0.4258	0.5596	9.6773
ViewCrafter [89]	202.93	0.4675	0.4830	11.0729
FlexWorld [18]	<u>183.69</u>	<u>0.5434</u>	<u>0.4111</u>	<u>12.7678</u>
PE-Field [7]	185.76	0.5301	0.4634	12.2003
Ours	**172.63**	**0.5546**	**0.3735**	**12.7902**

Fig. 4: Qualitative comparison under the **limited camera motion** setting. Our method maintains stable spatial layouts and scene structural consistency across views.

Fig. 5: Our approach models continuous camera motion characteristics. Sequences are shown from left to right: the input image (blue), intermediate frames reflecting the trajectory (red), and the final novel view (green).

duplicated structures, distorted geometric relationships, and locally incoherent content, especially when the camera motion becomes extensive. In contrast, our method better maintains geometric consistency across views and produces more natural and coherent novel-view results, effectively alleviating structural degradation and visual inconsistencies.

In addition, Fig. 6 shows qualitative results on the MannequinChallenge [45] dataset, which mainly focuses on human-centric scenes. Compared with other methods, our approach exhibits more stable identity preservation across views and thus produces more reliable results under camera motion.

Intermediate Trajectory Visualization. Furthermore, to provide deeper insights into our generation process, we visualize the intermediate synthesized frames along the camera trajectory in Fig. 5. This visualization demonstrates how our model smoothly and accurately models the continuous geometric transformations dictated by the camera motion. By maintaining structural coherence throughout the intermediate process, our approach aligns with the camera motion characteristics, ensuring precision in the final rendered views.

5.3 Robustness Analysis

Robustness to Point Cloud Quality Our framework remains robust even when the point clouds reconstructed by VGGT are noisy or incomplete. Since

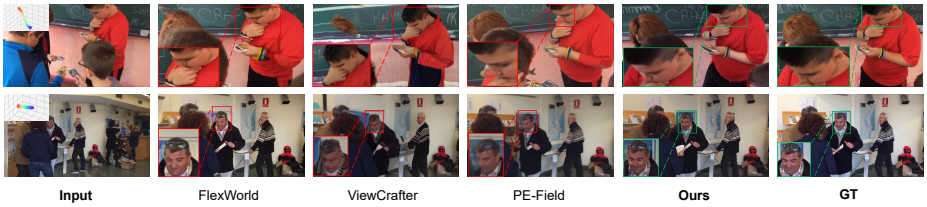

Fig. 6: Qualitative comparison on the **MannequinChallenge** dataset. Under camera motion, our method achieves more stable identity preservation compared with other methods, maintaining more consistent appearance.

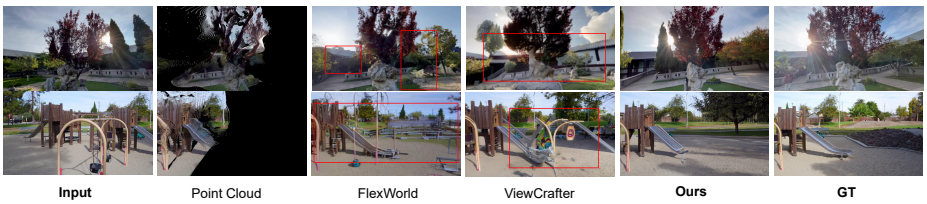

Fig. 7: Robustness to imperfect point clouds. Even with noisy and incomplete VGGT reconstruction, our method maintains consistent geometry and appearance.

the proposed soft contextual query mechanism aggregates geometric information through global attention rather than enforcing strict point-wise correspondence, imperfect point clouds are not directly mapped to generated pixels and their influence can be gradually corrected during denoising. As illustrated in Figure 7, despite noticeable noise and large missing regions in the reconstructed geometry, our method still produces visually consistent novel views with coherent structures and appearance, demonstrating strong robustness to imperfect geometric priors.

Robustness to Real-World Content Our framework also generalizes well to real-world images containing dynamic or non-rigid content. Although VGGT may reconstruct point clouds on regions with motion or deformation (e.g., waving flags or human faces), our soft geometric constraints prevent these imperfect priors from introducing severe artifacts. As shown in Figure 8, the generated views preserve structural consistency and appearance fidelity even in the presence of dynamic or non-rigid regions, producing coherent and realistic results.

5.4 Ablation Study

To verify the effectiveness of the proposed designs, we conduct comprehensive ablation studies on three key components: Frame-Decoupled Point Cloud Injection (FDPCI), Geometric Anchor Attention (GAA), and Trajectory-Endpoint Geometric Supervision (TEGS). All experiments are conducted on the DL3DV [48] dataset under extensive and limited camera motion settings. Due to space constraints, additional ablation studies are provided in the supplementary material.

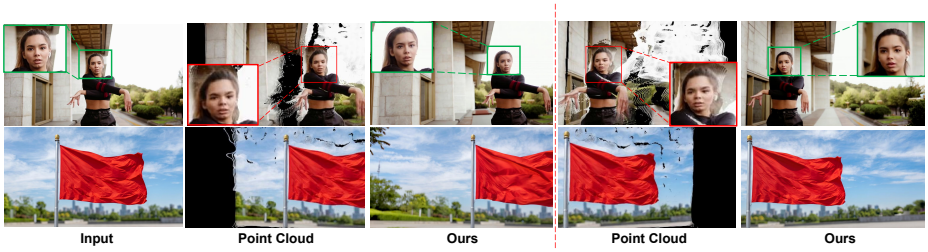

Fig. 8: Robustness on real-world scenes containing dynamic or non-rigid content.

Table 5: Ablation study on the **DL3DV dataset** under extensive and limited camera motion settings. We evaluate Frame-Decoupled Point Cloud Injection (FDPCI), Geometric Anchor Attention (GAA) and Trajectory-Endpoint Geometric Supervision (TEGS). The best results are highlighted in **bold**.

Setting	Extensive Camera Motion				Limited Camera Motion			
	FID↓	SSIM↑	LPIPS↓	PSNR↑	FID↓	SSIM↑	LPIPS↓	PSNR↑
w/o FDPCI	121.52	0.4270	0.3491	12.9989	75.79	0.4941	0.2218	15.7771
w/o GAA	115.83	0.4683	0.3284	13.3682	70.44	0.5060	0.2088	15.9162
w/o TEGS	119.43	0.4588	0.3379	13.0054	69.15	0.5050	0.2122	15.7987
Ours	**113.11**	**0.4830**	**0.3248**	**13.6067**	**69.05**	**0.5271**	**0.2065**	**16.3740**

Frame-Decoupled Point Cloud Injection. We first evaluate the effect of injecting point cloud sequences aligned with the target camera trajectory along the frame dimension as geometric priors. As shown in Table 5, removing FDPCI leads to a significant performance drop (e.g., LPIPS increases by 0.02 on average and SSIM drops by 0.06 in extensive motions), indicating its crucial role in maintaining structural consistency and perceptual quality. Qualitatively, as shown in Fig. 9, this approach effectively avoids object duplication and positional errors, maintaining more stable geometry under viewpoint changes.

Geometric Anchor Attention. We then evaluate the effect of Geometric Anchor Attention. As demonstrated in Table 5, the removal of GAA results in noticeable degradation across all metrics, proving that introducing the first-frame geometric features as anchors explicitly aligns cross-view features and preserves geometric structure. To further explore its optimal setting, we conduct a hyper-parameter analysis on the attention weight α, detailed in Table 6. Consistently applying the anchor attention improves generation quality, achieving the best performance at a balanced weight of $\alpha = 1.0$. Setting α too low ($\alpha = 0.1$) weakens the geometric alignment, while setting it too high ($\alpha = 1.5$) overly constrains the features, leading to slight performance drops.

Trajectory-Endpoint Geometric Supervision. Finally, we investigate the design of the Trajectory-Endpoint Geometric Supervision (TEGS). Incorporating TEGS significantly improves structural fidelity at target views and enhances geometric consistency compared to the baseline without it (w/o TEGS in Table 5). To determine the optimal supervision intensity, we further evaluate different

Table 6: Analysis on the weight α of Geometric Anchor Attention (GAA) under extensive and limited camera motion settings on the **DL3DV dataset**.

Setting	Extensive Camera Motion				Limited Camera Motion			
	FID↓	SSIM↑	LPIPS↓	PSNR↑	FID↓	SSIM↑	LPIPS↓	PSNR↑
FDPCI + TEGS + GAA ($\alpha = 0.1$)	115.17	0.4476	0.3343	13.2395	70.02	0.5096	0.2115	16.2663
FDPCI + TEGS + GAA ($\alpha = 0.5$)	113.42	0.4636	0.3311	13.3267	69.48	0.5084	0.2066	15.9713
FDPCI + TEGS + GAA ($\alpha = 1.0$) (ours)	**113.11**	**0.4830**	**0.3248**	**13.6067**	**69.05**	**0.5271**	**0.2065**	**16.3740**
FDPCI + TEGS + GAA ($\alpha = 1.5$)	114.80	0.4718	0.3252	13.6044	70.88	0.5103	0.2108	16.0774

Table 7: Analysis on the weight γ of Trajectory-Endpoint Geometric Supervision (TEGS) under extensive and limited camera motion settings on the **DL3DV dataset**.

Setting	Extensive Camera Motion				Limited Camera Motion			
	FID↓	SSIM↑	LPIPS↓	PSNR↑	FID↓	SSIM↑	LPIPS↓	PSNR↑
FDPCI + GAA + TEGS ($\gamma = 0.001$)	119.58	0.4394	0.3484	13.0486	70.98	0.4883	0.2327	15.5676
FDPCI + GAA + TEGS ($\gamma = 0.01$) (ours)	**113.11**	**0.4830**	**0.3248**	**13.6067**	**69.05**	**0.5271**	**0.2065**	**16.3740**
FDPCI + GAA + TEGS ($\gamma = 0.1$)	115.60	0.4616	0.3331	13.2588	71.57	0.4993	0.2169	15.8824

Table 8: Inference cost and quality comparison.

Method	Time (s) ↓	Memory (GB) ↓	PSNR ↑	LPIPS ↓
CameraCtrl [30]	45	25	10.3463	0.4537
MotionCtrl [73]	42	20	10.5097	0.5039
ViewCrafter [89]	140	24	13.9124	0.4088
FlexWorld [18]	240	27	<u>14.3485</u>	<u>0.3277</u>
PE-Field [7]	70	38	13.4911	0.3919
Ours	75	29	**14.9904**	**0.2657**

hyperparameter settings for TEGS in Table 7. The results show that the best geometric consistency and visual quality are achieved under our configuration.

Additionally, we conduct a qualitative control experiment where geometric supervision is applied *only* at trajectory endpoints, leaving intermediate frames completely unconstrained (denoted as "w/o intermediate supervision (w/o IS)" in Fig. 9). As demonstrated, this extreme setting produces substantially blurrier results, indicating that completely omitting geometric guidance on intermediate frames weakens the video model's inherent temporal continuity priors, thereby reducing the geometric stability of the final generated sequence.

5.5 Limitation

While our method demonstrates strong performance in camera-controllable view synthesis, two main limitations remain: (1) **Complex scenes and extreme viewpoint changes** (in Fig. 10) : When handling highly complex scenes or excessively large viewpoint variations, especially the latter, the introduced geometric references may become unreliable, leading to degraded geometric accuracy in the generated results; (2) **Inference efficiency**: Even with sparse temporal sampling to reduce the number of frames processed during inference, a certain number of

Fig. 9: Qualitative results of the ablation study. Without point cloud or intermediate supervision, the generated results suffer from object duplication.

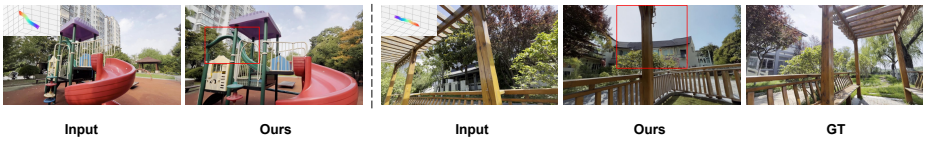

Fig. 10: Failure cases: Left—complex objects challenge geometry and texture preservation; Right—extreme camera changes impede geometric consistency.

frames still need to be generated. While this is more efficient than video generation models, the inference time is still slightly longer compared to single-frame image diffusion models. Using a LoRA [19] for accelerated sampling within the video model can further improve efficiency. To further clarify efficiency and the inference cost, we report results in Tab. 8. Despite using multi-frame latent generation, our method remains much more efficient than full video-generation approaches while achieving better quality; compared with lightweight baselines, it incurs only moderate additional runtime but significantly improves geometric consistency, thus achieving a favorable efficiency-quality trade-off.

6 Conclusion

We propose **UniGeo**, a camera-controllable view synthesis framework that leverages the prior of video diffusion models to enforce unified geometric guidance throughout the generation process. By systematically integrating geometric guidance across representation, architecture, and loss function, UniGeo overcomes the limitations of fragmented geometric injection, establishing reliable cross-view correspondences while ensuring structural integrity. Experiments demonstrate that UniGeo consistently outperforms existing methods in both geometric reliability and visual quality, providing a principled and effective solution for high-fidelity camera-controllable view synthesis across diverse camera motions.

Acknowledgements

This work was supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM103), the Fundamental Research Funds for the Central Universities (226-2025-00080), and the Earth System Big Data Platform of the School of Earth Sciences, Zhejiang University.

References

1. Albergo, M.S., Vanden-Eijnden, E.: Building normalizing flows with stochastic interpolants. arXiv preprint arXiv:2209.15571 (2022)
2. Ali, A., Bai, J., Bala, M., Balaji, Y., Blakeman, A., Cai, T., Cao, J., Cao, T., Cha, E., Chao, Y.W., et al.: World simulation with video foundation models for physical ai. arXiv preprint arXiv:2511.00062 (2025)
3. Avrahami, O., Patashnik, O., Fried, O., Nemchinov, E., Aberman, K., Lischinski, D., Cohen-Or, D.: Stable flow: Vital layers for training-free image editing. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 7877–7888 (June 2025)
4. Bahmani, S., Skorokhodov, I., Qian, G., Siarohin, A., Menapace, W., Tagliasacchi, A., Lindell, D.B., Tulyakov, S.: Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22875–22889 (2025)
5. Bahmani, S., Skorokhodov, I., Siarohin, A., Menapace, W., Qian, G., Vasilkovsky, M., Lee, H.Y., Wang, C., Zou, J., Tagliasacchi, A., et al.: Vd3d: Taming large video diffusion transformers for 3d camera control. arXiv preprint arXiv:2407.12781 (2024)
6. Bai, J., Xia, M., Fu, X., Wang, X., Mu, L., Cao, J., Liu, Z., Hu, H., Bai, X., Wan, P., et al.: Recammaster: Camera-controlled generative rendering from a single video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14834–14844 (2025)
7. Bai, Y., Li, H., Huang, Q.: Positional encoding field. arXiv preprint arXiv:2510.20385 (2025)
8. Bar-Tal, O., Ofri-Amar, D., Fridman, R., Kasten, Y., Dekel, T.: Text2live: Text-driven layered image and video editing. In: European conference on computer vision. pp. 707–723. Springer (2022)
9. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence **35**(8), 1798–1828 (2013)
10. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
11. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22563–22575 (2023)
12. Brack, M., Friedrich, F., Kornmeier, K., Tsaban, L., Schramowski, P., Kersting, K., Passos, A.: Ledits++: Limitless image editing using text-to-image models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8861–8870 (2024)
13. 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)
14. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., et al.: Video generation models as world simulators. OpenAI Blog **1**(8), 1 (2024)

15. 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)
16. Chan, E.R., Nagano, K., Chan, M.A., Bergman, A.W., Park, J.J., Levy, A., Aittala, M., De Mello, S., Karras, T., Wetzstein, G.: Generative novel view synthesis with 3d-aware diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4217–4229 (2023)
17. Chen, J., Xue, L., Xu, Z., Pan, X., Yang, S., Qin, C., Yan, A., Zhou, H., Chen, Z., Huang, L., et al.: Blip3o-next: Next frontier of native image generation. arXiv preprint arXiv:2510.15857 (2025)
18. Chen, L., Zhou, Z., Zhao, M., Wang, Y., Zhang, G., Huang, W., Sun, H., Wen, J.R., Li, C.: Flexworld: Progressively expanding 3d scenes for flexiable-view synthesis. arXiv preprint arXiv:2503.13265 (2025)
19. Contributors, L.: Lightx2v: Light video generation inference framework. `https://github.com/ModelTC/lightx2v` (2025)
20. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. arXiv preprint arXiv:2210.11427 (2022)
21. Cui, Y., Chen, H., Deng, H., Huang, X., Li, X., Liu, J., Liu, Y., Luo, Z., Wang, J., Wang, W., Wang, Y., Wang, C., Zhang, F., Zhao, Y., Pan, T., Li, X., Hao, Z., Ma, W., Chen, Z., Ao, Y., Huang, T., Wang, Z., Wang, X.: Emu3.5: Native multimodal models are world learners (2025), `https://arxiv.org/abs/2510.26583`
22. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., Shi, G., Fan, H.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
23. Domingos, P.: A few useful things to know about machine learning. Communications of the ACM **55**(10), 78–87 (2012)
24. Dong, H., Wang, W., Li, C., Lin, D.: Wan-alpha: High-quality text-to-video generation with alpha channel. arXiv e-prints pp. arXiv–2509 (2025)
25. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
26. Gao, R., Holynski, A., Henzler, P., Brussee, A., Martin-Brualla, R., Srinivasan, P., Barron, J.T., Poole, B.: Cat3d: Create anything in 3d with multi-view diffusion models. arXiv preprint arXiv:2405.10314 (2024)
27. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016), `http://www.deeplearningbook.org`
28. Guo, Y., Yang, C., Rao, A., Agrawala, M., Lin, D., Dai, B.: Sparsectrl: Adding sparse controls to text-to-video diffusion models. In: European Conference on Computer Vision. pp. 330–348. Springer (2024)
29. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
30. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for text-to-video generation. arXiv preprint arXiv:2404.02101 (2024)
31. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control.(2022). URL https://arxiv. org/abs/2208.01626 **3**, 3 (2022)

32. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
33. Hou, C., Chen, Z.: Training-free camera control for video generation. arXiv preprint arXiv:2406.10126 (2024)
34. Huang, Y., Huang, J., Liu, Y., Yan, M., Lv, J., Liu, J., Xiong, W., Zhang, H., Cao, L., Chen, S.: Diffusion model-based image editing: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
35. Huang, Y., Xie, L., Wang, X., Yuan, Z., Cun, X., Ge, Y., Zhou, J., Dong, C., Huang, R., Zhang, R., 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. pp. 8362–8371 (2024)
36. Ju, X., Zeng, A., Bian, Y., Liu, S., Xu, Q.: Pnp inversion: Boosting diffusion-based editing with 3 lines of code. In: The Twelfth International Conference on Learning Representations (2023)
37. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6007–6017 (2023)
38. Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Transactions on Graphics (ToG) **36**(4), 1–13 (2017)
39. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
40. Kuang, Z., Cai, S., He, H., Xu, Y., Li, H., Guibas, L.J., Wetzstein, G.: Collaborative video diffusion: Consistent multi-video generation with camera control. Advances in Neural Information Processing Systems **37**, 16240–16271 (2024)
41. Kulikov, V., Kleiner, M., Huberman-Spiegelglas, I., Michaeli, T.: Flowedit: Inversion-free text-based editing using pre-trained flow models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19721–19730 (2025)
42. Labs, B.F.: Flux. `https://github.com/black-forest-labs/flux` (2024)
43. Lai, W.S., Huang, J.B., Wang, O., Shechtman, E., Yumer, E., Yang, M.H.: Learning blind video temporal consistency. In: European Conference on Computer Vision (2018)
44. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)
45. Li, Z., Dekel, T., Cole, F., Tucker, R., Snavely, N., Liu, C., Freeman, W.T.: Learning the depths of moving people by watching frozen people. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4521–4530 (2019)
46. Li, Z., Liu, Z., Zhang, Q., Lin, B., Wu, F., Yuan, S., Yan, Z., Ye, Y., Yu, W., Niu, Y., et al.: Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. arXiv preprint arXiv:2510.16888 (2025)
47. Lin, B., Li, Z., Cheng, X., Niu, Y., Ye, Y., He, X., Yuan, S., Yu, W., Wang, S., Ge, Y., et al.: Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation. arXiv preprint arXiv:2506.03147 (2025)
48. Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., et al.: Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d

vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22160–22169 (2024)
49. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
50. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9298–9309 (2023)
51. Liu, S., Han, Y., Xing, P., Yin, F., Wang, R., Cheng, W., Liao, J., Wang, Y., Fu, H., Han, C., et al.: Step1x-edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761 (2025)
52. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
53. Ma, B., Gao, H., Deng, H., Luo, Z., Huang, T., Tang, L., Wang, X.: You see it, you got it: Learning 3d creation on pose-free videos at scale. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2016–2029 (2025)
54. 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)
55. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6038–6047 (2023)
56. Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. In: ACM SIGGRAPH 2023 conference proceedings. pp. 1–11 (2023)
57. Peng, X., Zheng, Z., Shen, C., Young, T., Guo, X., Wang, B., Xu, H., Liu, H., Jiang, M., Li, W., Wang, Y., Ye, A., Ren, G., Ma, Q., Liang, W., Lian, X., Wu, X., Zhong, Y., Li, Z., Gong, C., Lei, G., Cheng, L., Zhang, L., Li, M., Zhang, R., Hu, S., Huang, S., Wang, X., Zhao, Y., Wang, Y., Wei, Z., You, Y.: Open-sora 2.0: Training a commercial-level video generation model in $200k. arXiv preprint arXiv:2503.09642 (2025)
58. Ren, X., Shen, T., Huang, J., Ling, H., Lu, Y., Nimier-David, M., Müller, T., Keller, A., Fidler, S., Gao, J.: Gen3c: 3d-informed world-consistent video generation with precise camera control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6121–6132 (2025)
59. Rotstein, N., Yona, G., Silver, D., Velich, R., Bensaïd, D., Kimmel, R.: Pathways on the image manifold: Image editing via video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7857–7866 (2025)
60. Rout, L., Chen, Y., Ruiz, N., Caramanis, C., Shakkottai, S., Chu, W.S.: Semantic image inversion and editing using rectified stochastic differential equations. arXiv preprint arXiv:2410.10792 (2024)
61. Seo, J., Fukuda, K., Shibuya, T., Narihira, T., Murata, N., Hu, S., Lai, C.H., Kim, S., Mitsufuji, Y.: Genwarp: Single image to novel views with semantic-preserving generative warping. Advances in Neural Information Processing Systems **37**, 80220–80243 (2024)
62. Shi, R., Chen, H., Zhang, Z., Liu, M., Xu, C., Wei, X., Chen, L., Zeng, C., Su, H.: Zero123++: a single image to consistent multi-view diffusion base model. arXiv preprint arXiv:2310.15110 (2023)
63. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)

64. Song, W., Jiang, H., Yang, Z., Quan, R., Yang, Y.: Insert anything: Image insertion via in-context editing in dit. arXiv preprint arXiv:2504.15009 (2025)
65. Sun, W., Chen, S., Liu, F., Chen, Z., Duan, Y., Zhang, J., Wang, Y.: Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. arXiv preprint arXiv:2411.04928 (2024)
66. Van Hoorick, B., Wu, R., Ozguroglu, E., Sargent, K., Liu, R., Tokmakov, P., Dave, A., Zheng, C., Vondrick, C.: Generative camera dolly: Extreme monocular dynamic novel view synthesis. In: European Conference on Computer Vision. pp. 313–331. Springer (2024)
67. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
68. Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 **3**(4), 6 (2025)
69. Wang, G.H., Zhao, S., Zhang, X., Cao, L., Zhan, P., Duan, L., Lu, S., Fu, M., Zhao, J., Li, Y., Chen, Q.G.: Ovis-u1 technical report. arXiv preprint arXiv:2506.23044 (2025)
70. 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)
71. 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)
72. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
73. Wang, Z., Yuan, Z., Wang, X., Li, Y., Chen, T., Xia, M., Luo, P., Shan, Y.: Motionctrl: A unified and flexible motion controller for video generation. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024)
74. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report (2025), https://arxiv.org/abs/2508.02324
75. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., et al.: Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871 (2025)
76. Wu, J.Z., Ren, X., Shen, T., Cao, T., He, K., Lu, Y., Gao, R., Xie, E., Lan, S., Alvarez, J.M., et al.: Chronoedit: Towards temporal reasoning for image editing and world simulation. arXiv preprint arXiv:2510.04290 (2025)
77. Wu, P., Zhu, K., Liu, Y., Zhao, L., Zhai, W., Cao, Y., Zha, Z.J.: Improved video vae for latent video diffusion model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18124–18133 (2025)

78. Wu, R., Mildenhall, B., Henzler, P., Park, K., Gao, R., Watson, D., Srinivasan, P.P., Verbin, D., Barron, J.T., Poole, B., et al.: Reconfusion: 3d reconstruction with diffusion priors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21551–21561 (2024)
79. Xiao, S., Wang, Y., Zhou, J., Yuan, H., Xing, X., Yan, R., Li, C., Wang, S., Huang, T., Liu, Z.: Omnigen: Unified image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13294–13304 (2025)
80. Xu, D., Nie, W., Liu, C., Liu, S., Kautz, J., Wang, Z., Vahdat, A.: Camco: Camera-controllable 3d-consistent image-to-video generation. arXiv preprint arXiv:2406.02509 (2024)
81. Xu, R., Zhou, D., Shen, X., Ma, F., Yang, Y.: Phyedit: Towards real-world object manipulation via physically-grounded image editing. arXiv preprint arXiv:2604.07230 (2026)
82. Xu, S., Huang, Y., Pan, J., Ma, Z., Chai, J.: Inversion-free image editing with language-guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9452–9461 (2024)
83. Yang, S., Hou, L., Huang, H., Ma, C., Wan, P., Zhang, D., Chen, X., Liao, J.: Direct-a-video: Customized video generation with user-directed camera movement and object motion. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–12 (2024)
84. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
85. Yang, Z., Chen, G., Li, X., Wang, W., Yang, Y.: Doraemongpt: Toward understanding dynamic scenes with large language models (exemplified as a video agent). arXiv preprint arXiv:2401.08392 (2024)
86. Yoon, S.H., Li, M., Beaudouin, G., Wen, C., Azhar, M.R., Wang, M.: Splitflow: Flow decomposition for inversion-free text-to-image editing. arXiv preprint arXiv:2510.25970 (2025)
87. You, M., Zhu, Z., Liu, H., Hou, J.: Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. arXiv preprint arXiv:2405.15364 (2024)
88. Yu, Q., Chow, W., Yue, Z., Pan, K., Wu, Y., Wan, X., Li, J., Tang, S., Zhang, H., Zhuang, Y.: Anyedit: Mastering unified high-quality image editing for any idea. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26125–26135 (2025)
89. Yu, W., Xing, J., Yuan, L., Hu, W., Li, X., Huang, Z., Gao, X., Wong, T.T., Shan, Y., Tian, Y.: Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. arXiv preprint arXiv:2409.02048 (2024)
90. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. Advances in Neural Information Processing Systems **36**, 31428–31449 (2023)
91. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023)
92. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
93. Zhang, Z., Chen, Z., Yang, Z., Yang, Y.: Are image-to-video models good zero-shot image editors? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2090–2103 (2026)
94. Zhang, Z., Xie, J., Lu, Y., Yang, Z., Yang, Y.: In-context edit: Enabling instructional image editing with in-context generation in large-scale diffusion transform-

ers. In: Advances in Neural Information Processing Systems (NeurIPS) (2025), arXiv:2504.20690
95. Zhao, H., Ma, X.S., Chen, L., Si, S., Wu, R., An, K., Yu, P., Zhang, M., Li, Q., Chang, B.: Ultraedit: Instruction-based fine-grained image editing at scale. Advances in Neural Information Processing Systems **37**, 3058–3093 (2024)
96. Zheng, G., Li, T., Jiang, R., Lu, Y., Wu, T., Li, X.: Cami2v: Camera-controlled image-to-video diffusion model. arXiv preprint arXiv:2410.15957 (2024)
97. Zheng, S., Peng, Z., Zhou, Y., Zhu, Y., Xu, H., Huang, X., Fu, Y.: Vidcraft3: Camera, object, and lighting control for image-to-video generation. arXiv preprint arXiv:2502.07531 (2025)
98. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404 (2024)
99. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. arXiv preprint arXiv:1805.09817 (2018)
100. Zhou, Z., Ma, F., Gui, C., Xia, X., Fan, H., Yang, Y., Chua, T.S.: Anchorflow: Training-free 3d editing via latent anchor-aligned flows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14387–14397 (2026)