DualCamCtrl: Dual-Branch Diffusion Model for Geometry-Aware Camera-Controlled Video Generation

Hongfei Zhang[1⋆], Kanghao Chen[1,5⋆], Zixin Zhang[1,5], Harold H. Chen[1,5], Yuanhuiyi Lyu[1], Kun Zhou[4], Yuqi Zhang[3], Shuai Yang[1], and Ying-cong Chen[1,2⋆⋆]

[1]HKUST (GZ) [2]HKUST [3]Fudan University [4]Shenzhen University [5]Knowin
hongfeifayezhang@gmail.com, kchen879@connect.hkust-gz.edu.cn
yingcongchen@ust.hk

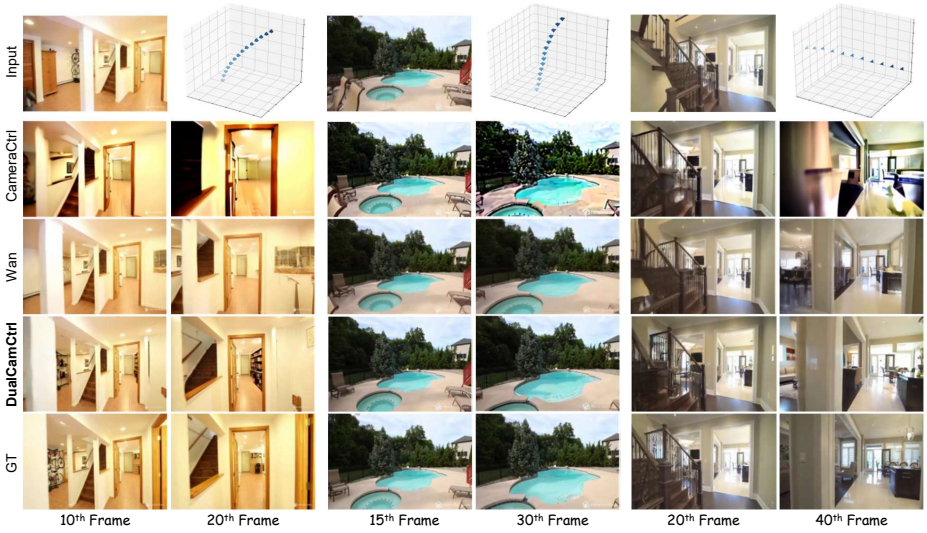

Fig. 1: Comparison with state-of-the-art methods [23, 78] on camera-controlled video generation. Under an identical target camera trajectory and a single input image, our approach achieves the closest adherence to the camera motion and yields the best perceptual quality.

Abstract. This paper presents **DualCamCtrl**, a novel architecture for camera-controlled video generation that internalizes geometric reasoning into the diffusion process. Existing methods rely solely on ray-based camera conditioning, which entangles appearance and geometry modeling within a single representation, or resort to external 3D cache and multi-stage pipelines that are hard to jointly optimize. DualCamCtrl addresses this by generating depth as a co-evolving modality alongside RGB through a dual-branch framework, enabling geometry and appearance to inform each other throughout denoising. To govern this cross-modal interaction, we propose the **S**emant**I**c **G**uided **M**utual **A**lignment

⋆ Equal contribution.
⋆⋆ Corresponding author.

(SIGMA) mechanism, which schedules RGB–depth fusion in a stage-aware manner based on our empirical finding that depth and camera poses exert asymmetric influence across the denoising trajectory. Together with a two-stage training strategy, these designs enable DualCamCtrl to disentangle and jointly evolve appearance and geometry, generating videos that faithfully adhere to specified camera trajectories. Extensive experiments demonstrate that DualCamCtrl achieves more consistent camera-controlled video generation **with over 40% reduction** on camera rotation errors compared with prior methods. Project Page: https://soyouthinkyoucantell.github.io/dualcamctrl-page.

1 Introduction

Advances in video diffusion models [1, 6, 18–21, 26, 27, 34, 41, 46, 81, 83, 85, 95, 97] have significantly improved the quality of video generation from textual descriptions. These developments have extended beyond basic text-to-video (T2V) or image-to-video (I2V) generation, with growing efforts toward integrating control mechanisms into the video generation process, such as camera control [1, 23, 24, 28, 36, 42, 86, 99–101], motion transfer [11, 31, 54, 57, 62, 84, 99], object manipulation [16, 37, 61, 66, 75, 88], and multi-modality prediction [14, 30, 58, 68, 69, 82, 92–94]. Among them, camera control has emerged as a particularly promising direction, enabling natural camera movements and viewpoint transitions that bridge the gap between generative modeling and real-world cinematography, supporting applications from virtual directors to interactive 3D scene generation [6, 12, 19, 32, 83, 89, 101].

Building on this progress, camera-conditioned video diffusion models [1, 23, 24, 86, 101] have been explored to follow camera trajectories during video synthesis. However, most approaches rely solely on camera poses or their ray-conditioned representations (*i.e.*, Plücker embeddings) as control signals.

Although ray-conditioned models [1, 23, 32, 86, 101] effectively encode camera motion through per-pixel ray directions, they force the network to simultaneously learn *what* to generate (appearance) and *where* to place it in 3D space (geometry), using only ray-based signals that carry no explicit scene structure. We argue that this **entanglement of appearance and geometric reasoning** is a fundamental bottleneck in camera-controlled video generation: the model must implicitly infer scene geometry from appearance cues alone, inevitably leading to suboptimal camera motion consistency. Meanwhile, some recent methods [49, 64, 98] attempt to address this by leveraging external or multi-stage 3D reconstruction pipelines (*e.g.*, point cloud caching and re-projection). While these approaches decouple geometry from appearance to some extent, they introduce complex multi-step workflows and cannot jointly optimize geometry and appearance within a unified generation process. These observations motivate a fundamentally different design: an architecture that *internalizes* geometric reasoning into the video generation process itself, enabling geometry and appearance to jointly evolve under camera control. Depth is a natural candidate for

such geometric grounding, yet our systematic study reveals that *naive integration strategies consistently fail*: conditioning on a single-frame depth map [10] lacks temporal context and leads to unnatural scene structures, while jointly modeling RGB and depth within a single branch introduces modality interference that degrades synthesis quality. The challenge thus lies not in *whether* to incorporate depth, but in *how* to architect the cross-modal interaction so that geometric and appearance representations can coexist and mutually reinforce each other without interference.

To this end, we propose **DualCamCtrl**, a novel camera-control architecture that internalizes geometric reasoning through parallel depth generation. Unlike prior methods that depend on external control signals at every frame or multi-stage 3D reconstruction pipelines, DualCamCtrl generates depth as a co-evolving modality within the latent diffusion process itself. Conditioned on shared camera poses, the RGB and depth branches jointly denoise, allowing geometry and appearance to mutually inform each other throughout generation—a capability absent in architectures where 3D information is externally computed and injected as a static signal.

We further introduce the **S**emant**I**c **G**uided **M**utual **A**lignment (SIGMA), a novel mechanism to govern cross-modal interaction between the two branches. SIGMA is motivated by our empirical analysis, which reveals that depth and camera poses exert *asymmetric influence* on the RGB denoising trajectory: camera poses dominate global structure formation in early denoising stages, while depth progressively refines geometric consistency in later stages. Accordingly, SIGMA prioritizes semantic fidelity in the early phase and activates geometry-guided mutual feedback in the later phase, enabling the two branches to co-evolve in a principled, stage-aware manner. Additionally, a two-stage training strategy is adopted to stabilize the learning process, consisting of a decoupled stage followed by a fusion stage.

To summarize, our main contributions are threefold:

1. We identify the entanglement of appearance and geometric reasoning as a core bottleneck in camera-controlled video generation, and systematically demonstrate that naive depth integration strategies fail to resolve it, revealing that the key challenge lies in *how* to architect the cross-modal interaction.
2. We propose **DualCamCtrl**, a new architecture that internalizes geometric reasoning via parallel depth generation. Guided by our empirical analysis of modality-specific denoising dynamics, we design the SIGMA mechanism for stage-aware cross-modal alignment and a two-stage training strategy, enabling geometry and appearance to jointly evolve without heavy external 3D pipelining.
3. Extensive experiments validate the effectiveness of our approach, achieving more than 40% reductions in rotation errors compared with previous SOTA methods, while preserving high-fidelity visual quality.

2 Related Work

Foundation model of video generation. Recent foundational diffusion-based video generation models, such as CogVideoX [96], Stable Video Diffusion (SVD) [5], Wan [78], Lumiere [3], VideoCrafter [8,9], ModelScope-T2V [80], Latte [52], and Open-Sora [43,60], have established powerful baselines for text-to-video and conditional video synthesis. These models employ large-scale latent diffusion architectures to achieve high temporal consistency and fine-grained scene dynamics. Building upon these foundational models, subsequent studies have explored their adaptation to various downstream video generation tasks, including camera control, [1, 7, 23, 24, 28, 36, 42, 50, 86, 99–101], motion transfer [11,31,54,57,62,84,99], object manipulation [16,37,61,66,75,88], and multi-modality prediction [14, 30, 58, 68, 69, 82, 92–94]. In this work, we focus on the problem of camera-controllable video generation, aiming to enable more consistent camera control video generation built upon diffusion-based foundation models [78].

Camera control for video models. Early efforts toward camera-controllable video generation were pioneered by methods such as MotionCtrl [86], which explicitly encode camera poses to guide video diffusion models. Following this line, CameraCtrl [23] and VD3D [2] further introduce ray-based conditioning (*i.e.*, Plücker embeddings [71]) to inject camera parameters into pretrained diffusion backbones. More recently, CameraCtrl-II [24] and Wan [78] extend this idea by incorporating camera pose information through pre-DiT conditioning. Meanwhile, AC3D [1] provides a deeper analysis of Video DiT behavior and mitigates training data limitations by constructing dynamic-scene datasets with rigid camera motion. Despite these advances, existing approaches still struggle to preserve the intended camera trajectory.

Geometry-aware video generation. Incorporating explicit 3D geometric cues into video generation has attracted growing attention. Traditional approaches based on neural radiance fields [56, 72] and 3D Gaussian splatting [38] achieve geometry-consistent novel view synthesis but require dense posed images and per-scene optimization. By leveraging generative priors, ReconFusion [90] and CAT3D [17] enable sparse-view reconstruction, yet still rely on per-scene optimization, limiting their scalability. More recent works such as ReconX [49], ViewCrafter [98], and GEN3C [64] incorporate depth and point clouds to decouple geometry from appearance and enable camera-controlled generation from sparse inputs. However, these approaches rely on disjointed, multi-stage pipelines that explicitly warp or project 3D representations—a paradigm that enforces rigid geometry and is inherently tailored for novel view synthesis. In contrast, our method internalizes geometric reasoning directly into an end-to-end latent diffusion process, enabling the joint evolution of appearance and geometry while preserving the flexibility of dynamic generative priors.

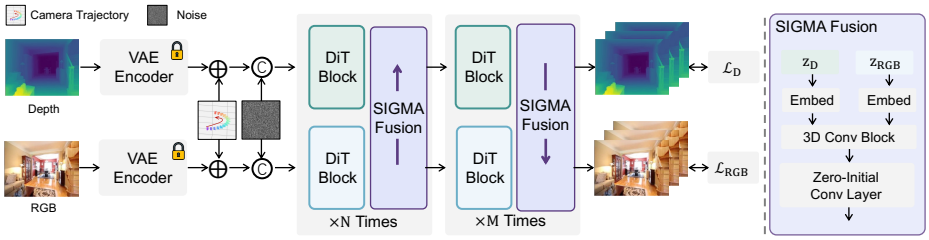

Fig. 2: Overall architecture of DualCamCtrl. DualCamCtrl adopts a dual-branch framework that simultaneously generates RGB and depth video latents from an input image and its corresponding depth map. The two latents are then element-wise added to the encoded Plücker embedding and concatenated with noise (Sec. 3.2). Subsequently, the two modalities interact through our proposed SIGMA mechanism and fusion block (Sec. 3.3). During training, both predictions are supervised by their respective loss functions (Sec. 3.4).

3 Method

In this section, we introduce the *DualCamCtrl* framework, which enables more geometry-aware camera-conditioned video diffusion by integrating geometric cues (*i.e.*, depth) in a dual-branch manner. We first provide necessary preliminaries in Sec. 3.1. In Sec. 3.2, we describe how depth is injected via our dual-branch architecture. We introduce the **S**emant**I**c **G**uided **M**utual **A**lignment and 3D fusion strategy for interaction between RGB and depth branches in Sec. 3.3. Sec. 3.4 details our two-stage training pipeline.

3.1 Preliminary

The video diffusion process typically involves a gradual denoising procedure, where a noisy latent representation of the video z_t is progressively transformed into a video sequence. In camera-controlled video generation, the process is mainly conditioned on two inputs: the context condition c (*i.e.* images or textual descriptions), and the camera pose $(\mathbf{R}, \mathbf{t})$. Camera positions are encoded through ray-based conditioning [1, 23, 78], where each pixel is associated with a Plücker ray representation derived from $(\mathbf{R}, \mathbf{t})$.

Formally, given a video $\mathbf{V} \in \mathbb{R}^{T\times 3\times H\times W}$, where T is the number of frames, and H, W are the height and width of each frame. This video is first encoded into a latent space using a pretrained encoder ε, producing $z_0 = \varepsilon(\mathbf{V})$, where $z_0 \in \mathbb{R}^{T'\times C'\times h\times w}$. Under our setting [78], the latent z has temporal length $T' = \frac{T-1}{4} + 1$, channel dimension $C' = 16$ and spatial dimensions $h = H/8$, $w = W/8$. The diffusion process then gradually denoises z_t at each timestep t, conditioned on inputs c, which can be an image c_{image} or text c_{text}, as well as camera pose parameters $(\mathbf{R}, \mathbf{t})$. The training objective is to predict the noise ϵ added to the latent (Eq. (1)):

$$\mathcal{L} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\left[\left\|\epsilon - \theta\big(z_t, t, c, \mathbf{R}, \mathbf{t}\big)\right\|^2\right], \tag{1}$$

where θ denotes the denoising network. During the inference stage, the denoising process is performed iteratively over a sequence of timesteps t. After denoising in the latent space, the output $\hat{z}$ is decoded by the decoder D to obtain the predicted video $\hat{\mathbf{V}}$: $\hat{\mathbf{V}} = D(\hat{z})$.

3.2 Dual-Branch Camera-Controlled Video Generation

Depth has long been recognized as a crucial cue for scene understanding as it explicitly encodes scene geometry and spatial layout [13, 22, 30, 44, 45, 53, 65, 69, 73, 87]. However, despite the importance of geometric priors, the integration of depth cues in camera-conditioned video generation remains unexplored. In this section, we investigate how depth information can be leveraged to enhance camera-conditioned video synthesis. Specifically, under the I2V setting, the input consists of a single reference frame, and a monocular depth estimator [10] allows us to predict a plausible depth map for that frame. However, relying solely on this single-frame depth is insufficient for guiding long-range video generation, as it provides only a static snapshot of the scene's geometry without maintaining temporal consistency. Moreover, naively coupling the generation of RGB and depth modalities tends to cause interference between them. This motivates us to design a framework that can propagate and effectively utilize depth information throughout the video sequence, thereby assisting RGB synthesis while minimizing interference. To this end, we propose a dual-branch architecture, where RGB and depth features are processed in parallel, each conditioned on the same camera poses to generate modality-specific representations (Fig. 2). This design allows the model to produce coherent information for both modalities under consistent camera motion.

Formally, given an input image $\mathbf{I}_{\text{RGB}} \in \mathbb{R}^{3\times H\times W}$, we first employ a monocular depth estimator [10] to obtain its depth map $\mathbf{I}_{\text{D}} \in \mathbb{R}^{1\times H\times W}$. Following previous works [25, 30, 68, 93], the estimated depth is replicated along the channel dimension to match the channel number of the RGB input. Both the RGB image and the replicated depth are then encoded into the latent space using a pretrained encoder ε, producing their respective latent representations $z^{\mathbf{I}_{\text{RGB}}}$ and $z^{\mathbf{I}_{\text{D}}}$. The two latent representations are both zero-padded in the frame channel to match the length of noise. Then they are element-wise added with the same encoded plücker embedding which represents the camera pose $(\mathbf{R}, \mathbf{t})$. Afterwards, they are concatenated with the same Gaussian noise ϵ and processed by a denoising network θ (*i.e.* Diffusion Transformer [59]), which consists of two parallel branches corresponding to the RGB and depth modalities. The model θ jointly predicts the denoised latent sequences, yielding an RGB video latent $\hat{z}^{\mathbf{V}_{\text{RGB}}}$ and a depth video latent $\hat{z}^{\mathbf{V}_{\text{D}}}$.

During training, the predicted latents are supervised by the ground-truth video latents $z^{\mathbf{V}_{\text{RGB}}}$ and $z^{\mathbf{V}_{\text{D}}}$, which are obtained by encoding the ground-truth RGB-D video $\mathbf{V}^{\text{RGB}}$ and $\mathbf{V}^{\text{D}}$. During inference, a pretrained decoder $\boldsymbol{D}$ is employed to reconstruct the corresponding RGB from the predicted latents. The training strategy is further described in Sec. 3.4. Details of the T2V setting are provided in the supplementary material due to space limitations (Fig. 2).

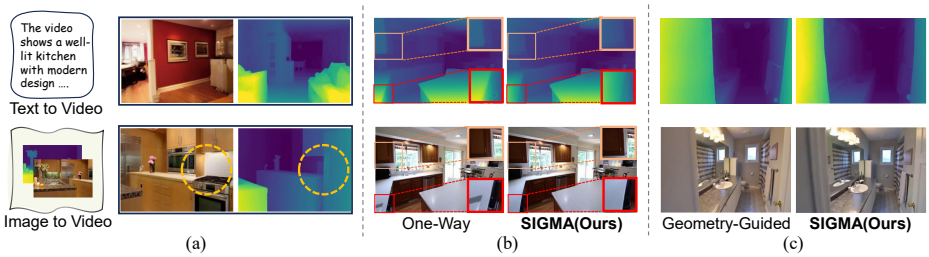

Fig. 3: (a) Illustration of modality misalignment. Independent RGB and depth latent evolution leads misalignment across frames. This motivates the design of our SIGMA strategy to establish coherent cross-modal alignment. **(b) Comparison with one-way alignment.** One-way alignment transfers information unidirectionally, leading to misalignment on local semantics. **(c) Comparison with geometry-guided alignment** Under the geometry-guided setting, geometry cues evolved too quickly and become inconsistent with RGB motion.

3.3 Injecting Depth Cues for Mutual Alignment

Semantic Guided Mutual Alignment. Based on our dual-branch framework, one branch synthesizes the RGB sequence, while the other generates the corresponding depth sequence. Despite being conditioned on the same input (*i.e.* image or text), the two branches evolve independently, resulting in inconsistency between their outputs, which necessitates an effective alignment strategy (Fig. 3). To address this, we first performed a CKA analysis between the dynamic and ground truth depth latent features, as well as the camera pose embeddings (Fig. 4). Our findings suggest that early layers in the RGB branch are more closely related to the camera pose. Building on this, we explored further and discovered that simple fusion strategies remain insufficient. We observe that *one-way alignment*, which transfers features from one modality to another without feedback, fails to preserve semantic consistency. On the other hand, *geometry-guided alignment* overemphasizes geometric cues, disrupting appearance coherence. In response, we introduce **S**emant**I**c **G**uided **M**utual **A**lignment (SIGMA). Motivated by this observation, SIGMA adopts a semantic-guided bidirectional design: in the early layers, RGB features anchor semantic structure, while later layers incorporate depth feedback to refine geometry. We empirically validate this design choice in Tab. 4.

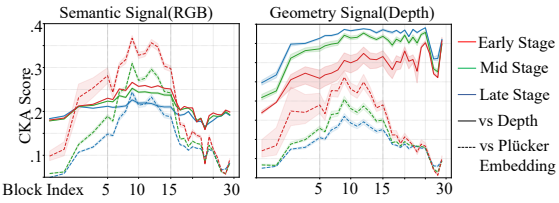

Fig. 4: We conduct linear CKA analysis [40] on the model obtained after the decoupled training stage (Sec. 3.4), computed over 500 randomly sampled validation sequences from RealEstate10K [102] at early, mid, and late denoising stages. Shaded regions indicate ±1 standard deviation over 3 random seeds. At early layers, RGB features are more closely related to the Plücker embedding, suggesting a closer association with camera pose across all denoising stages.

Notably, this design is motivated by two empirical observations. **Observa-**

tion.❶ Priority matters: Appearance should first dominate the initial stages to ensure semantic fidelity, while depth is introduced later as a complementary corrective signal that refines geometric consistency. **Observation.❷ Mutual feedback matters:** Enabling both branches to inform each other avoids the imbalance of one-way alignment, leading to more consistent representation. Together, these principles enable SIGMA to more effectively harmonize semantic and geometric representations, ultimately yielding more visually coherent video generation (Tab. 4).

3D Fusion Strategy. Our experiments reveal that achieving temporal and spatial consistency between RGB and depth sequences cannot rely solely on alignment strategies (Fig. 5). While SIGMA effectively coordinates the two branches, conventional shallow projection layers (*e.g.*, linear feature injection as in [19, 83, 100]) still ignore the temporal ordering and spatial layout of video, treating all latent pixels uniformly across frames and often producing temporally inconsistent outputs. To effectively capture spatiotemporal cues, we introduce 3D convolutions into the fusion block, enabling the model to fuse semantic and geometry information over both spatial and temporal dimensions (Fig. 2). To balance computational cost and parameter efficiency, we further adopt an efficient bottleneck design inspired by [29, 67, 74]. A frame-wise gating mechanism is involved to adaptively modulate the fusion strength according to temporal dynamics.

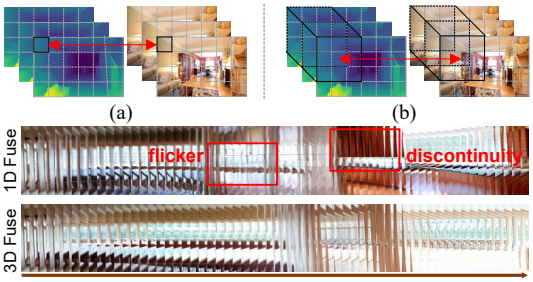

Fig. 5: Comparison of fusion strategies. (Upper.a) Previous shallow linear fusion operates in a pixel-wise manner, ignoring temporal and spatial context and causing inconsistency over time. (Upper.b) We introduce 3D Fusion strategy to extend the fusion formulation from 1D to 3D by incorporating 3D operations. (Lower) Traditional linear-layer fusion often leads to visual artifacts, while our methods produce smoother, more coherent results.

3.4 Two-stage Training Pipeline

Recall that our training objective is twofold: to enable each modality to develop generative competence and to foster effective cross-modal interaction. We empirically found that the depth branch, which starts from scratch, is initially interpreted by the model as producing low-light effects when generating content (Fig. 6). However, directly training both RGB and depth branches jointly leads to severe convergence issues: the depth stream fails to learn meaningful geometry, which destabilizes overall optimization (Fig. 9). To address this, we adopt a two-stage training strategy. In the first, decoupling stage, appearance and geometry are learned independently; in the second, fusion stage, cross-branch interaction is introduced via a fusion block. The detailed procedure is described below.

Fig. 6: Effect of staged training. *(Left)*: Feeding depth directly into the diffusion model does not cause immediate collapse, but the model interprets it as a low-light scenario for generation. *(Middle)*: Without decoupling appearance and geometry during training, the model initially runs but ultimately fails to converge. *(Right)*: As a result, the generated motion is inconsistent with the input geometry.

Initialization and Decoupled Stage. To incorporate knowledge from pre-trained diffusion models, we first initialize both the RGB and depth branches with the same model weights [78]. The primary goal of this stage is to enable the model to learn a robust depth synthesis module that can be used as a complementary signal for the RGB branch. However, no existing dataset provides simultaneously scene-level realistic video motion, corresponding camera parameters, and ground-truth depth [4,15,35,39,47,55,91]. To address this, we employ state-of-the-art monocular depth estimation methods to predict depth maps for all frames and use them as supervision for the depth branch [10]. To avoid interference between the two modalities, we do not perform any cross-branch fusion at this stage. This separation ensures that each branch can independently capture its respective cues, appearance for RGB and geometry for depth, without destabilizing the training process (Fig. 9). Interestingly, we observed that although the depth branch cannot generate the target sequence initially, it still interprets depth as a hazy image and uses this as a condition for generation (Fig. 4). Furthermore, considering that the depth supervision is derived from pseudo labels, we incorporate additional robustness strategies during training to mitigate noise and further improve the quality of the learned depth representations (Sec.4).

Fusion Stage. In the second stage, we enable fusion between the RGB and depth branches to exploit their complementary strengths: the RGB branch provides rich appearance cues, while the depth branch conveys geometric structure. The training objective remains a combination of depth and RGB losses (Eq. (4)). We zero-initialize the fusion block, allowing their influence to emerge gradually.

Loss Function. Let $\gamma \in \{0, 1\}$ indicate whether cross-branch fusion is enabled, where $\gamma = 0$ corresponds to the decoupled stage and $\gamma = 1$ corresponds to the fusion stage. We define the 3D-aware cross features $h_t^{\mathrm{RGB}\to\mathrm{D}}$ (from RGB to Depth) and $h_t^{\mathrm{D}\to\mathrm{RGB}}$ (from Depth to RGB), with the corresponding losses for each branch as follows (Eqs. (2) and (3)):

$$\mathcal{L}_{\mathrm{RGB}} = \mathbb{E}\left[\left\| v_t^{\mathrm{RGB}} - \theta_{\mathrm{RGB}}(z_t^{\mathrm{RGB}}, t, c, \mathbf{R}, \mathbf{t}; \gamma\, h_t^{\mathrm{D}\to\mathrm{RGB}}) \right\|^2\right], \tag{2}$$

$$\mathcal{L}_{\mathrm{D}} = \mathbb{E}\left[\left\| v_t^{\mathrm{D}} - \theta_{\mathrm{D}}\left(z_t^{\mathrm{D}}, t, c, \mathbf{R}, \mathbf{t}; \gamma\, h_t^{\mathrm{RGB}\to\mathrm{D}}\right) \right\|^2\right]. \tag{3}$$

The overall loss is then defined as (Eq. (4)):

$$\mathcal{L}_{\mathrm{Overall}} = \mathcal{L}_{\mathrm{RGB}} + \lambda \mathcal{L}_{\mathrm{D}}. \tag{4}$$

Table 1: Quantitative comparisons on I2V setting. ↑ / ↓ denotes higher/lower is better. We highlight the best and second best.

Method	RealEstate10K							DL3DV						
	FVD ↓	FID ↓	CLIPSIM ↑	FC ↑	MS ↑	RE ↓	TE ↓	FVD ↓	FID ↓	CLIPSIM ↑	FC ↑	MS ↑	RE ↓	TE ↓
MotionCtrl [86]	137.4	71.70	0.2401	0.9621	0.9733	2.80	1.04	158.5	98.51	0.2009	0.9386	0.9360	1.10	1.73
CameraCtrl [23]	118.7	69.90	0.2358	0.9623	0.9860	2.38	1.03	144.4	85.37	0.2015	0.9569	0.9745	1.04	1.71
Seva [101]	104.2	76.69	0.3059	0.9570	0.9815	2.62	0.32	206.0	130.9	0.2044	0.9443	0.9471	1.06	1.67
Wan [78]	109.2	77.80	0.2859	0.9567	0.9863	2.08	0.32	127.9	70.77	0.2022	0.9512	0.9731	1.01	1.64
DualCamCtrl	80.38	49.85	0.2998	0.9677	0.9886	1.25	0.23	92.2	53.89	0.2047	0.9637	0.9763	0.88	1.39

Fig. 7: Comparison between our method and other state-of-the-art approaches. Given the same camera pose and input image as generation conditions, our method achieves the best alignment between camera motion and scene dynamics, producing the most visually accurate video. The '+' signs marked in the figure serve as **anchors to indicate specific reference points for better visual comparison**.

4 Experiments

4.1 Experimental Setup

Baselines. We evaluate our method against several state-of-the-art camera-controlled video generation models. In the I2V setting, we compare our approach with MotionCtrl [86] and CameraCtrl [24], as well as two recent strong baselines, Wan-V2.1 [78] and SEVA [101]. In the T2V setting, we compare with CameraCtrl, MotionCtrl, and the most recent AC3D [1]. All baseline models are evaluated under the same split and evaluation metrics for a consistent comparison.

Implementation Details. We use Wan-V2.1 as the backbone video diffusion model. Following [1,24,86,101], we train our model on the RealEstate10K [102]. It contains over 60 M video clips with per-frame camera parameters. Training is conducted on 8 H100 GPUs. We follow the original training scheme of the baseline model with 1000 denoising steps, using rectified flow [48, 51] as the training target.

Depth Modeling for Robust Control. As discussed in Sec. 3.4, we supervise depth prediction using pseudo-labeled depth maps as ground truth [10], which inevitably contain failure cases. To enhance robustness against imperfect depth estimates and to promote joint learning beyond our dual-branch design and SIGMA, we early-stop the decoupled training stage to capture only coarse geometric cues, and randomly drop the depth loss to prevent overfitting to noisy depth signals. We put more analysis in the appendix.

Table 2: Quantitative comparisons on T2V setting across RealEstate10K and DL3DV.

Method	RealEstate10K						DL3DV					
	FVD ↓	CLIPSIM ↑	FC ↑	MS ↑	RE ↓	TE ↓	FVD ↓	CLIPSIM ↑	FC ↑	MS ↑	RE ↓	TE ↓
MotionCtrl [86]	506.9	0.2744	0.9471	0.9734	2.90	1.06	746.7	0.2033	0.9448	0.9462	1.17	3.23
CameraCtrl [23]	426.8	0.2734	0.9526	0.9750	2.68	0.98	675.5	0.2027	0.9518	0.9610	1.01	4.00
AC3D [1]	415.6	0.3044	0.9496	0.9811	2.67	0.38	452.8	0.2078	0.9529	0.9869	1.06	2.11
DualCamCtrl	408.1	0.3154	0.9540	0.9918	1.23	0.25	427.4	0.2090	0.9544	0.9807	0.83	1.53

Evaluation Metrics. We evaluate video generation performance using widely adopted metrics, including Fréchet Video Distance (FVD) [77], Fréchet Inception Distance (FID) [76], and CLIP similarity (CLIPSim) based on CLIP [63]. We further assess Frame Consistency (FC) and Motion Strength (MS) using VBench [33]. Rotation and translation errors (RE, TE) are computed with fine-tracking from VGGSfM [79], excluding tracking failures ($\approx 5\%$) and near-static cases. We also conduct a human study with over 30 participants for qualitative evaluation (Fig. 10).

4.2 Main Results

Quantitative Results. We compare our proposed DualCamCtrl against recent state-of-the-art video generation methods, including MotionCtrl [86], CameraCtrl [23], Seva [101], Wan [78], and AC3D [1]. Our evaluations on the RealEstate10K and DL3DV datasets under both Image-to-Video (I2V) and Text-to-Video (T2V) settings, with results summarized in Tables 1 and 2. As shown, DualCamCtrl achieves the best overall performance across nearly all metrics. We highlight three key advantages of our approach:

1. **State-of-the-art Generative Quality:** Our method establishes a new state-of-the-art in distribution-level realism. In the I2V setting, DualCamCtrl significantly reduces the FVD from 104.2 to 80.38 on RealEstate10K, and from 127.9 to 92.2 on DL3DV, demonstrating clearly superior visual fidelity compared to existing baselines.
2. **Precise Camera Controllability:** The core strength of DualCamCtrl lies in its geometric accuracy. We observe a substantial reduction in both Rotation Error (RE) and Translation Error (TE). Specifically, in the T2V setting on RealEstate10K, our method decreases RE by over 40% (from 2.67 to 1.23) and TE from 0.38 to 0.25. This precision in motion modeling mitigates structural warping, directly contributing to the improved FVD scores.
3. **Spatiotemporal Consistency:** Beyond geometric precision, DualCamCtrl maintains leading scores in CLIPSIM, Frame Consistency (FC), and Motion Smoothness (MS). These results confirm that our approach effectively balances fine-grained camera control with high semantic alignment and temporal coherence across diverse datasets and generation settings.

Qualitative Results. As illustrated in Figs. 7 and 8, DualCamCtrl synthesizes videos that better adhere to the intended camera motions while achieving

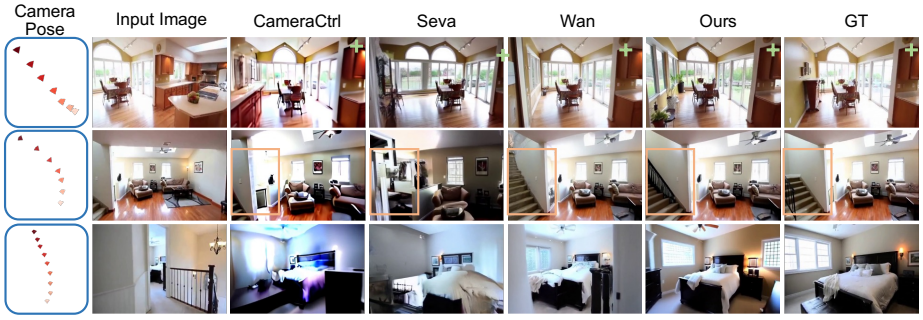

Fig. 8: Comparison under noticeable camera motion. Under noticeable camera motion, a substantial portion of the scene must be synthesized from previously unobserved regions. Our method preserves geometric consistency while achieving the highest visual fidelity. The '+' symbols indicate reference points for precise visual comparison.

noticeably higher visual fidelity and stronger temporal coherence in the generated regions compared to baseline methods. Furthermore, our approach attains the highest overall preference rate in the user study (Fig. 10). These results demonstrate that incorporating depth-aware geometric modeling not only improves quantitative performance but also produces outputs that are subjectively preferred by human evaluators.

4.3 Ablation Study

In this section, we conduct ablation studies to evaluate the contributions of key components in our **DualCamCtrl** framework on the RealEstate10K dataset.

Effect of Incorporating Depth. To assess the benefit of depth guidance, we compare our model with and without depth guidance. As shown in Fig. 9, removing depth leads to a notable degradation in FVD and MS, indicating weaker temporal stability and geometry understanding. This confirms that depth provides strong structural cues for camera-controlled scene generation. Here, we control the parameters to isolate the effect of depth.

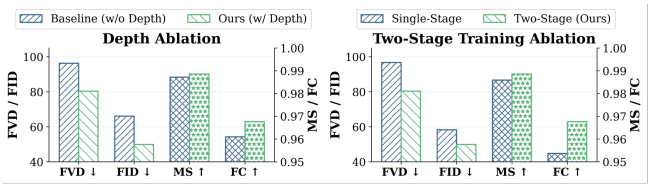

Fig. 9: Ablation studies on our method. Impact of **depth conditioning** and **two-stage training** on performance. Metrics include FVD (↓), FID (↓), MS (↑), and FC (↑). Both design choices improve video generation quality.

Effect of Two-Stage Training. We further evaluate the effectiveness of our two-stage training scheme. As seen in Fig. 9, direct joint training (only fusion stage) leads to unstable convergence and suboptimal temporal coherence. The

Table 3: Ablation on 3D fusion block.

Variant	FVD ↓	FID ↓	MS ↑	FC ↑
ControlNet-Style (1D)	112.3	61.8	0.9819	0.9641
Spatial (2D) Fusion	98.4	60.1	0.9815	0.9643
Spatial Temporal (3D) Fusion	85.4	52.7	0.9861	0.9648
3D Fusion + Frame Gating	80.4	49.9	0.9886	0.9677

two-stage variant achieves superior results across all metrics, suggesting that decoupling appearance and geometry learning before fusion significantly stabilizes training and enhances cross-modal consistency.

Effect of 3D fusion strategy. We perform a progressive ablation to evaluate each component of the 3D fusion block. As shown in Tab. 3, adding spatial, 3D spatiotemporal, and frame-gating modules progressively improves performance.

Effect of Semantic Guided Mutual Alignment. We further investigate the impact of cross-branch communication and fusion depth. Three fusion strategies are compared: (1) *One-way Alignment*, (2) *Geometry-Guided Alignment*, (3) *Semantic Guided Mutual Alignment (Ours)*. In Tab. 4, the left two columns indicate the layer ranges where RGB and depth features are injected, allowing us to analyze the effect of fusion depth under each strategy. As shown in Tab. 4, SIGMA consistently achieves the best video quality metrics as well as camera motion consistency, indicating that semantic guided mutual alignment is more effective than one-way or geometry-guided conditioning.

Table 4: Analysis on alignment mechanisms.

#RGB	#Depth	FVD ↓	FID ↓	MS ↑	FC ↑	RE ↓	TE ↓
-	1-5	91.4	59.1	0.984	0.966	1.76	0.36
-	1-10	87.0	58.4	0.984	0.964	1.56	0.37
-	1-15	83.2	59.7	0.985	0.964	1.55	0.35
6-10	1-5	87.4	54.2	0.985	0.966	1.56	0.33
6-15	1-5	84.6	53.2	0.987	0.966	1.55	0.34
11-15	1-10	84.2	53.7	0.987	0.965	1.55	0.34
1-5	6-10	80.9	49.5	0.987	0.965	1.27	0.25
1-5	6-15	80.4	49.9	0.989	0.968	1.25	0.25
1-10	11-15	80.3	50.2	0.989	0.969	1.27	0.23

Robustness to Depth Quality. We evaluate the robustness of DualCamCtrl under varying depth quality by introducing realistic degradations to the depth input, including spatial downsampling, low-resolution monocular estimation, and simulated sensor noise. As shown above, all degraded settings incur only marginal performance drops compared to the default configuration (e.g., Rot. Err +0.03∼+0.12). Importantly, camera motion accuracy and video quality remain substantially better than the baseline, indicating that our method primarily leverages coarse geometric structure rather than relying on precise pixel-level depth fidelity. These results demonstrate strong robustness to imperfect depth estimation in practical scenarios.

Table 5: Robustness to different depth quality.

Methods & Degradations	**Rot. Err ↓**	**Trans. Err ↓**	**FVD ↓**
Seva	2.62	0.32	104.2
Ours (Default)	**1.25**	**0.23**	**80.4**
Downsample 4×	1.28 (+0.03)	0.25 (+0.02)	82.9
MiDaS (Low Res)	1.28 (+0.03)	0.26 (+0.03)	82.1
Kinect Depth (Sim.)	1.37 (+0.12)	0.28 (+0.05)	84.7

Backbone Generality and Scalability. To further substantiate generality, we evaluate DualCamCtrl on diffusion backbones of different capacities. As shown above, our method consistently improves camera accuracy and video quality across both a weaker

Table 6: Ablation on model variants (Baseline / **Ours**).

Backbone	**Rot. Err ↓**	**Trans. Err ↓**	**FVD ↓**
CogVideo-2B (Weaker)	2.37 / **1.74**	0.49 / **0.33**	121.1 / **95.7**
CogVideo-5B (Stronger)	2.01 / **1.21**	0.37 / **0.27**	104.8 / **75.9**

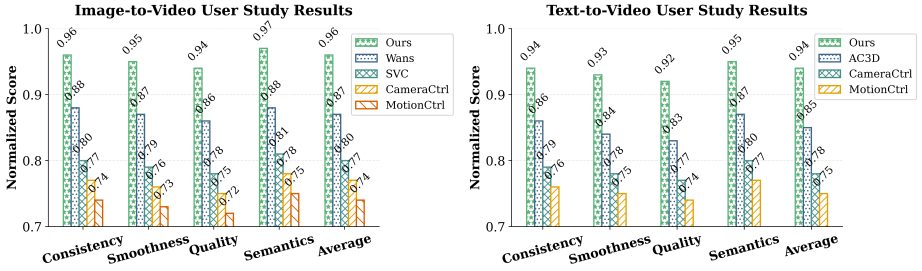

Fig. 10: User study results (The scores are 0–1 normalized to better illustrate the results, ↑ indicates better). **Please zoom in for better details.**

Fig. 11: Failure case under large motion. When the inter-frame motion is large, severe artifacts emerge.

(2B) and a stronger (5B) model [96]. Notably, larger backbones benefit even more from our design (e.g., Rot. Err 2.01→1.21; FVD 104.8 →75.9), demonstrating favorable scalability with model size. These results confirm that our framework is backbone-agnostic and can be seamlessly integrated into diffusion video models of varying scales.

User Study. Beyond objective metrics, we conduct a user study to assess perceptual video quality and controllability (Fig. 10). Participants evaluate four aspects: (1) *Camera Consistency*, (2) *Smoothness*, (3) *Visual Quality*, and (4) *Semantic Consistency*. As shown in Fig. 10, our method consistently outperforms all baselines in both image-to-video and text-to-video settings. Notably, improvements are most significant in camera consistency and semantic alignment, validating the effectiveness of our geometry-aware control design.

4.4 Discussion and Limitations

While our method performs strongly, we identify three limitations that suggest potential directions for future improvement. We discuss them below with the aim

Large Motion. Our method suffers from issues related to large motion, which is a common challenge in video generation tasks. Other methods use NVS to resolve ambiguities and mitigate flickering in video generation [17, 49, 64, 90, 98]. However, despite our advancements, large motion still presents a difficulty in stable and coherent end-to-end video generation (Fig. 11).

Model Capacity and Parameter Efficiency. Although our dual-branch architecture inherently increases the overall parameter count by duplicating the backbone for separate RGB and depth processing, our final model size (**2.7B**) remains comparable to prior designs [23, 101] (e.g., CameraCtrl at 2.4B). This efficiency is primarily attributed to the compact nature of the underlying WanV2.1 foundation model (1.3B). While deploying two parallel branches is acceptable and yields clear performance advantages at this scale, this full duplication may become less favorable when scaling to substantially larger video diffusion models. A promising future direction is to develop more parameter-efficient depth guidance, for instance, by distilling a lightweight depth branch that can effectively condition the RGB pathway while keeping the overall model footprint minimal.

Table 7: Comparison of model capacity, trainable parameters, and inference latency. All are tested on a $14 \times 576 \times 320$ shape.

Method	Total Params.	Latency
CameraCtrl	2.4B	40s
SEVA	1.5B	>2min
Ours	**2.7B**	**49s**

5 Conclusion

We presented DualCamCtrl, a dual-branch video diffusion model that integrates depth for more accurate camera-controlled video generation. By introducing the SemantIc Guided Mutual Alignment (SIGMA) mechanism and a two-stage training process, **DualCamCtrl** effectively synchronizes RGB and depth sequences, improving geometry awareness. Our experiments show over 40% reduction in rotation errors compared to previous methods, offering new insights into camera-controlled video generation and hopefully benefiting other video generation tasks.

References

1. 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) 2, 4, 5, 10, 11
2. 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) 4
3. Bar-Tal, O., Chefer, H., Tov, O., Michaeli, T., Dekel, T., et al.: Lumiere: A space-time diffusion model for video generation. arXiv preprint arXiv:2401.12945 (2024), https://arxiv.org/abs/2401.12945 4
4. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5855–5864 (2021) 9

5. Blattmann, A., Dockhorn, T., Kulal, S., Rombach, R., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023), https://arxiv.org/abs/2311.15127 4
6. Cai, S., Ceylan, D., Gadelha, M., Huang, C.H.P., Wang, T.Y., Wetzstein, G.: Generative rendering: Controllable 4d-guided video generation with 2d diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7611–7620 (2024) 2
7. Chen, D.Y., Guo, Y., Yang, S., Mu, T.J., Hu, S.M.: Beyond inpainting: Unleash 3d understanding for precise camera-controlled video generation. arXiv preprint arXiv:2601.10214 (2026) 4
8. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., et al.: Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512 (2023) 4
9. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024), https://openaccess.thecvf.com/content/CVPR2024/papers/Chen_VideoCrafter2_Overcoming_Data_Limitations_for_High-Quality_Video_Diffusion_Models_CVPR_2024_paper.pdf 4
10. Chen, S., Guo, H., Zhu, S., Zhang, F., Huang, Z., Feng, J., Kang, B.: Video depth anything: Consistent depth estimation for super-long videos. arXiv:2501.12375 (2025) 3, 6, 9, 10
11. Chen, W., Ji, Y., Wu, J., Wu, H., Xie, P., Li, J., Xia, X., Xiao, X., Lin, L.: Control-a-video: Controllable text-to-video diffusion models with motion prior and reward feedback learning. arXiv preprint arXiv:2305.13840 (2023) 2, 4
12. Chen, Y., Rao, A., Jiang, X., Xiao, S., Ma, R., Wang, Z., Xiong, H., Dai, B.: Cinepregen: Camera controllable video previsualization via engine-powered diffusion. arXiv preprint arXiv:2408.17424 (2024) 2
13. Choe, J., Im, S., Rameau, F., Kang, M., Kweon, I.S.: Volumefusion: Deep depth fusion for 3d scene reconstruction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16086–16095 (2021) 6
14. Dong, Y.J., Zhao, W., Xu, J., Shan, Y., Zhang, S.H.: Depthsync: Diffusion guidance-based depth synchronization for scale-and geometry-consistent video depth estimation. arXiv preprint arXiv:2507.01603 (2025) 2, 4
15. Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T.B., Vanhoucke, V.: Google scanned objects: A high-quality dataset of 3d scanned household items. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2553–2560. IEEE (2022) 9
16. Feng, R., Weng, W., Wang, Y., Yuan, Y., Bao, J., Luo, C., Chen, Z., Guo, B.: Ccedit: Creative and controllable video editing via diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6712–6722 (2024) 2, 4
17. 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) 4, 14
18. Geng, D., Herrmann, C., Hur, J., Cole, F., Zhang, S., Pfaff, T., Lopez-Guevara, T., Aytar, Y., Rubinstein, M., Sun, C., et al.: Motion prompting: Controlling video generation with motion trajectories. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1–12 (2025) 2

19. Gu, Z., Yan, R., Lu, J., Li, P., Dou, Z., Si, C., Dong, Z., Liu, Q., Lin, C., Liu, Z., Wang, W., Liu, Y.: Diffusion as shader: 3d-aware video diffusion for versatile video generation control. arXiv preprint arXiv:2501.03847 (2025) 2, 8
20. Guo, X., Zheng, M., Hou, L., Gao, Y., Deng, Y., Wan, P., Zhang, D., Liu, Y., Hu, W., Zha, Z., et al.: I2v-adapter: A general image-to-video adapter for diffusion models. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–12 (2024) 2
21. 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) 2
22. Gupta, K., Jabbireddy, S., Shah, K., Shrivastava, A., Zwicker, M.: Improved modeling of 3d shapes with multi-view depth maps. In: International Conference on 3D Vision (3DV) (2020) 6
23. 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) 1, 2, 4, 5, 10, 11, 15
24. He, H., Yang, C., Lin, S., Xu, Y., Wei, M., Gui, L., Zhao, Q., Wetzstein, G., Jiang, L., Li, H.: Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. arXiv preprint arXiv:2503.10592 (2025) 2, 4, 10
25. He, J., Li, H., Yin, W., Liang, Y., Li, L., Zhou, K., Liu, H., Liu, B., Chen, Y.C.: Lotus: Diffusion-based visual foundation model for high-quality dense prediction. arXiv preprint arXiv:2409.18124 (2024) 6
26. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022) 2
27. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in neural information processing systems **35**, 8633–8646 (2022) 2
28. Hou, C., Chen, Z.: Training-free camera control for video generation. arXiv preprint arXiv:2406.10126 (2024) 2, 4
29. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017) 8, 1
30. Hu, W., Gao, X., Li, X., Zhao, S., Cun, X., Zhang, Y., Quan, L., Shan, Y.: Depthcrafter: Generating consistent long depth sequences for open-world videos. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2005–2015 (2025) 2, 4, 6
31. Huang, C.P., Wu, Y.S., Chung, H.K., Chang, K.P., Yang, F.E., Wang, Y.C.F.: Videomage: Multi-subject and motion customization of text-to-video diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17603–17612 (2025) 2, 4
32. Huang, T., Zheng, W., Wang, T., Liu, Y., Wang, Z., Wu, J., Jiang, J., Li, H., Lau, R.W., Zuo, W., et al.: Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. arXiv preprint arXiv:2506.04225 (2025) 2
33. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024) 11, 4
34. Jain, Y., Nasery, A., Vineet, V., Behl, H.: Peekaboo: Interactive video generation via masked-diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8079–8088 (2024) 2

35. Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanæs, H.: Large scale multi-view stereopsis evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 406–413 (2014) 9
36. Jin, W., Dai, Q., Luo, C., Baek, S.H., Cho, S.: Flovd: Optical flow meets video diffusion model for enhanced camera-controlled video synthesis. arXiv preprint arXiv:2502.08244 (2025) 2, 4
37. Kahatapitiya, K., Karjauv, A., Abati, D., Porikli, F., Asano, Y.M., Habibian, A.: Object-centric diffusion for efficient video editing. In: European Conference on Computer Vision. pp. 91–108. Springer (2024) 2, 4
38. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023) 4
39. 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) 9
40. Kornblith, S., Norouzi, M., Lee, H., Hinton, G.: Similarity of neural network representations revisited. In: International conference on machine learning. pp. 3519–3529. PMlR (2019) 7
41. Koroglu, M., Caselles-Dupré, H., Jeanneret, G., Cord, M.: Onlyflow: Optical flow based motion conditioning for video diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6226–6236 (2025) 2
42. 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) 2, 4
43. Lab, P.Y., etc., T.A.: Open-sora-plan (Apr 2024). https://doi.org/10.5281/zenodo.10948109, https://doi.org/10.5281/zenodo.10948109 4
44. Laidlow, T., Czarnowski, J., Leutenegger, S.: Deepfusion: Real-time dense 3d reconstruction for monocular slam using single-view depth and gradient predictions. CoRR **abs/2207.12244** (2022), https://arxiv.org/abs/2207.12244 6
45. Lapid, A., Achituve, I., Bracha, L., Fetaya, E.: Gd-vdm: Generated depth for better diffusion-based video generation. arXiv preprint arXiv:2306.11173 (2023) 6
46. Liang, J., Fan, Y., Zhang, K., Timofte, R., Van Gool, L., Ranjan, R.: Movideo: Motion-aware video generation with diffusion model. In: European Conference on Computer Vision. pp. 56–74. Springer (2024) 2
47. 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) 9, 4
48. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022) 10
49. Liu, F., Sun, W., Wang, H., Wang, Y., Sun, H., Ye, J., Zhang, J., Duan, Y.: Reconx: Reconstruct any scene from sparse views with video diffusion model. arXiv preprint arXiv:2408.16767 (2024) 2, 4, 14
50. Liu, L., Li, W., Zhang, D., Wang, S., Jiao, S.: Idcnet: Guided video diffusion for metric-consistent rgbd scene generation with precise camera control. arXiv preprint arXiv:2508.04147 (2025) 4
51. 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) 10
52. Ma, X., Zhang, W., Zhou, P., et al.: Latte: Latent diffusion transformer for video generation. arXiv preprint arXiv:2401.03048 (2024), https://arxiv.org/abs/2401.03048 4

53. Menini, D., Kumar, S., Oswald, M.R., Sandström, E., Sminchisescu, C., Gool, L.V.: A real-time online learning framework for joint 3d reconstruction and semantic segmentation of indoor scenes. CoRR **abs/2108.05246** (2021), https://arxiv.org/abs/2108.05246 6
54. Meral, T.H.S., Yesiltepe, H., Dunlop, C., Yanardag, P.: Motionflow: Attention-driven motion transfer in video diffusion models. arXiv preprint arXiv:2412.05275 (2024) 2, 4
55. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. ACM Transactions on Graphics (ToG) **38**(4), 1–14 (2019) 9
56. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021) 4
57. Motamed, S., Van Gansbeke, W., Van Gool, L.: Investigating the effectiveness of cross-attention to unlock zero-shot editing of text-to-video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7406–7415 (2024) 2, 4
58. Pallotta, E., Azar, S.M., Li, S., Zatsarynna, O., Gall, J.: Syncvp: Joint diffusion for synchronous multi-modal video prediction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13787–13797 (2025) 2, 4
59. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023) 6
60. Peng, X., Zheng, Z., Shen, C., et al.: Open-sora 2.0: Training a commercial-level video generation model in $200k. arXiv preprint arXiv:2503.09642 (2025), https://arxiv.org/abs/2503.09642 4
61. Peruzzo, E., Goel, V., Xu, D., Xu, X., Jiang, Y., Wang, Z., Shi, H., Sebe, N.: Vase: Object-centric appearance and shape manipulation of real videos. arXiv preprint arXiv:2401.02473 (2024) 2, 4
62. Pondaven, A., Siarohin, A., Tulyakov, S., Torr, P., Pizzati, F.: Video motion transfer with diffusion transformers. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22911–22921 (2025) 2, 4
63. 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) 11
64. 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 Computer Vision and Pattern Recognition Conference. pp. 6121–6132 (2025) 2, 4, 14
65. Rosinol, A., Leonard, J.J., Carlone, L.: Probabilistic volumetric fusion for dense monocular slam. CoRR **abs/2210.01276** (2022), https://arxiv.org/abs/2210.01276 6
66. Saini, N., Bodla, N., Shrivastava, A., Ravichandran, A., Zhang, X., Shrivastava, A., Singh, B.: Invi: Object insertion in videos using off-the-shelf diffusion models. arXiv preprint arXiv:2407.10958 (2024) 2, 4
67. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018) 8, 1

68. Shao, J., Yang, Y., Zhou, H., Zhang, Y., Shen, Y., Guizilini, V., Wang, Y., Poggi, M., Liao, Y.: Learning temporally consistent video depth from video diffusion priors (2024), https://arxiv.org/abs/2406.01493 2, 4, 6
69. Shao, J., Yang, Y., Zhou, H., Zhang, Y., Shen, Y., Guizilini, V., Wang, Y., Poggi, M., Liao, Y.: Learning temporally consistent video depth from video diffusion priors. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22841–22852 (2025) 2, 4, 6
70. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1874–1883 (2016) 4
71. Sitzmann, V., Rezchikov, S., Freeman, B., Tenenbaum, J., Durand, F.: Light field networks: Neural scene representations with single-evaluation rendering. Advances in Neural Information Processing Systems **34**, 19313–19325 (2021) 4
72. Tancik, M., Weber, E., Ng, E., Li, R., Yi, B., Wang, T., Kristoffersen, A., Austin, J., Salahi, K., Ahuja, A., et al.: Nerfstudio: A modular framework for neural radiance field development. In: ACM SIGGRAPH 2023 conference proceedings. pp. 1–12 (2023) 4
73. Tateno, K., Tombari, F., Laina, I., Navab, N.: Cnn-slam: Real-time dense monocular slam with learned depth prediction. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 624–632 (2017). https://doi.org/10.1109/ICCV.2017.75 6
74. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018) 8, 1
75. Tu, S., Dai, Q., Cheng, Z.Q., Hu, H., Han, X., Wu, Z., Jiang, Y.G.: Motioneditor: Editing video motion via content-aware diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7882–7891 (2024) 2, 4
76. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018) 11
77. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Fvd: A new metric for video generation. arXiv preprint (2019) 11
78. 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) 1, 4, 5, 9, 10, 11, 2, 3
79. Wang, J., Karaev, N., Rupprecht, C., Novotny, D.: Vggsfm: Visual geometry grounded deep structure from motion (2023) 11
80. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023), https://arxiv.org/abs/2308.06571 4

81. Wang, L., Li, Y., Chen, Z., Wang, J.H., Zhang, Z., Zhang, H., Lin, Z., Chen, Y.C.: Transpixeler: Advancing text-to-video generation with transparency. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18229–18239 (2025) 2, 6
82. Wang, Q., Eldesokey, A., Mendiratta, M., Zhan, F., Kortylewski, A., Theobalt, C., Wonka, P.: Vidseg: Training-free video semantic segmentation based on diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22985–22994 (2025) 2, 4
83. Wang, Q., Luo, Y., Shi, X., Jia, X., Lu, H., Xue, T., Wang, X., Wan, P., Zhang, D., Gai, K.: Cinemaster: A 3d-aware and controllable framework for cinematic text-to-video generation. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–10 (2025) 2, 8
84. Wang, X., Li, X., Hu, Y., Zhu, H., Hou, C., Lan, C., Chen, Z.: Tiv-diffusion: Towards object-centric movement for text-driven image to video generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 7988–7996 (2025) 2, 4
85. Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., et al.: Lavie: High-quality video generation with cascaded latent diffusion models. International Journal of Computer Vision **133**(5), 3059–3078 (2025) 2
86. 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) 2, 4, 10, 11
87. Weder, S., Schönberger, J.L., Pollefeys, M., Oswald, M.R.: Routedfusion: Learning real-time depth map fusion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4886–4896 (2020). https://doi.org/10.1109/CVPR42600.2020.00494 6
88. Wu, B., Chuang, C.Y., Wang, X., Jia, Y., Krishnakumar, K., Xiao, T., Liang, F., Yu, L., Vajda, P.: Fairy: Fast parallelized instruction-guided video-to-video synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8261–8270 (2024) 2, 4
89. Wu, J.Z., Zhang, Y., Turki, H., Ren, X., Gao, J., Shou, M.Z., Fidler, S., Gojcic, Z., Ling, H.: Difix3d+: Improving 3d reconstructions with single-step diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26024–26035 (2025) 2
90. 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) 4, 14
91. Wu, T., Zhang, J., Fu, X., Wang, Y., Ren, J., Pan, L., Wu, W., Yang, L., Wang, J., Qian, C., et al.: Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 803–814 (2023) 9
92. Xi, D., Wang, J., Liang, Y., Qiu, X., Huo, Y., Wang, R., Zhang, C., Li, X.: Omnivdiff: Omni controllable video diffusion for generation and understanding. arXiv preprint arXiv:2504.10825 (2025) 2, 4
93. Yang, H., Huang, D., Yin, W., Shen, C., Liu, H., He, X., Lin, B., Ouyang, W., He, T.: Depth any video with scalable synthetic data. arXiv preprint arXiv:2410.10815 (2024) 2, 4, 6

94. Yang, L., Qi, L., Li, X., Li, S., Jampani, V., Yang, M.H.: Unified dense prediction of video diffusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28963–28973 (2025) 2, 4
95. 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) 2
96. Yang, Z., Teng, J., Zheng, W., Ding, M., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2025), https://arxiv.org/abs/2408.06072 4, 14
97. Yeh, Y.Y., Huang, J.B., Kim, C., Xiao, L., Nguyen-Phuoc, T., Khan, N., Zhang, C., Chandraker, M., Marshall, C.S., Dong, Z., et al.: Texturedreamer: Image-guided texture synthesis through geometry-aware diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4304–4314 (2024) 2
98. 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) 2, 4, 14
99. Zhao, R., Gu, Y., Wu, J.Z., Zhang, D.J., Liu, J.W., Wu, W., Keppo, J., Shou, M.Z.: Motiondirector: Motion customization of text-to-video diffusion models. In: European Conference on Computer Vision. pp. 273–290. Springer (2024) 2, 4
100. 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) 2, 4, 8
101. Zhou, J.J., Gao, H., Voleti, V., Vasishta, A., Yao, C.H., Boss, M., Torr, P., Rupprecht, C., Jampani, V.: Stable virtual camera: Generative view synthesis with diffusion models. arXiv preprint arXiv:2503.14489 (2025) 2, 4, 10, 11, 15
102. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. ACM Trans. Graph. (Proc. SIGGRAPH) **37** (2018), https://arxiv.org/abs/1805.09817 7, 10, 4