

OmniCamera: A Unified Framework for Multi-task Video Generation with Arbitrary Camera Control

Yukun Wang^{1,2*}, Ruihuang Li², Jiale Tao², Shiyuan Yang^{2,3}, Liyi Chen^{2,4}, Zhantao Yang², Handz², Yulan Guo¹, Shuai Shao², and Qinglin Lu²

¹Sun Yat-sen University ²Hunyuan, Tencent ³CityU ⁴PolyU

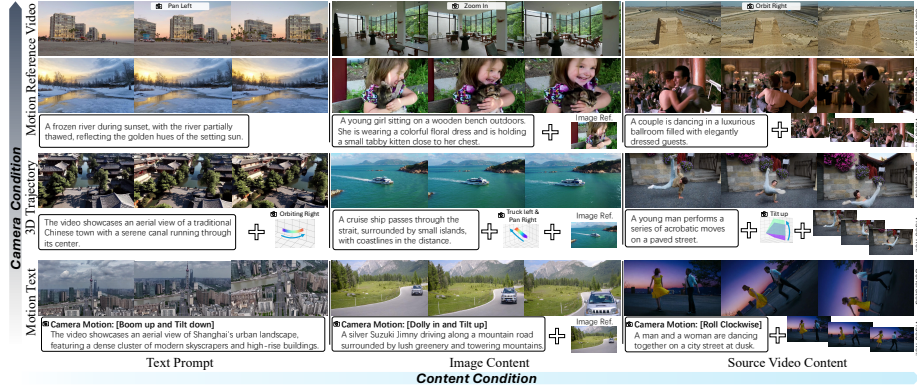


Fig. 1: We propose **OmniCamera**, a unified framework that conceptually decouples video generation into two independent control dimensions: camera pose and scene content. It seamlessly integrates three camera conditions (text, 3D trajectory, and motion reference video) with three content conditions (text prompt, image, and source video).

Abstract. Video fundamentally intertwines two crucial axes: the dynamic content of a scene and the camera motion through which it is observed. However, existing generation models often entangle these factors, limiting independent control. In this work, we introduce OmniCamera, a unified framework designed to explicitly disentangle and command these two dimensions. This compositional approach enables flexible video generation by allowing arbitrary pairings of camera and content conditions, unlocking unprecedented creative control. To overcome the fundamental challenges of modality conflict and data scarcity inherent in such a system, we present two key innovations. First, we construct OmniCAM, a novel hybrid dataset combining curated real-world videos with synthetic data that provides diverse paired examples for robust multi-task learning. Second, we propose a Dual-level Curriculum Co-Training strategy that mitigates modality interference and synergistically learns from diverse data sources. This strategy operates on two levels: first, it progressively introduces control modalities by difficulties (condition-level),

* Work done during an internship at Tencent Hunyuan.

and second, trains for precise control on synthetic data before adapting to real data for photorealism (data-level). As a result, OmniCamera achieves state-of-the-art performance, enabling flexible control for complex camera movements while maintaining superior visual quality.

1 Introduction

Fundamentally, every video is the physical projection of real-world scene content observed through continuous camera poses in 3D space. Driven by this universal mechanism, we propose to conceptually decouple video generation into two independent control dimensions: scene content and camera pose. While simulating realistic cinematographic operations is crucial for professional applications, current approaches typically restrict control to a single modality or focus on narrow tasks as shown in Table. 1. For instance, existing methods separately utilize textual descriptions [13, 15, 31], 3D trajectories [1, 2, 16, 22], or reference videos [26] for camera motion. They often struggle with the inherent limitations of each modality (e.g., text is too coarse, trajectories are hard to acquire) and fail to support the free combination of diverse content sources and camera conditions.

To address these limitations and theoretically encompass all conceivable forms of video generation, we introduce *OmniCamera*, a unified framework that seamlessly integrates these independent conditions. Specifically, as shown in Fig. 1, the **camera conditions** involve textual descriptions, explicit trajectory matrices, or reference motion videos, and **content conditions** can be provided by text prompts, images, or source videos. This framework offers three significant advantages. Firstly, our synergistic co-training strategy leverages the complementary strengths of diverse data sources. It learns precise camera control from the accurate geometry of synthetic data (e.g., UE5), while simultaneously learning photorealism from real-world videos. Secondly, as illustrated in Fig. 2, by flexibly combining multiple camera conditions, users can synthesize arbitrarily complex and nuanced camera movements. Thirdly, our model achieves remarkable parameter efficiency by employing a single set of weights to handle all nine distinct combinations, eliminating the need for specialized models.

Creating a system capable of unified multi-modal camera control is fraught with challenges, foremost among them being the inherent conflict between different control modalities. A naive joint training approach, which simply mixes data from various conditions (e.g., text, reference videos, trajectories), often leads to mutual interference. Besides, there is a profound scarcity of real-world data that simultaneously offers high visual quality and precise camera annotations, making it difficult to train a model that excels in both aspects.

To address these challenges, we present OmniCAM, the first hybrid dataset for camera control, comprising both real-world and synthetic videos and providing diverse types of paired data to facilitate multitask learning. As shown in Tab. 2, OmniCAM is the largest among publicly available datasets and uniquely supports multi-condition training across camera and content dimensions. The



Fig. 2: Compositional camera control with multi-modal conditions. Omni-Camera enables flexible combinations of multiple conditioning inputs, such as overlaying text-guided camera motion on top of trajectory or reference-video control, efficiently extending and diversifying camera motion effects.

real-world portion of the dataset is meticulously curated through a comprehensive pipeline—including trajectory estimation, filtering, classification, and matching—to select high-quality video pairs with reliable camera trajectories.

To leverage the unique structure of this dataset, we further propose a dual-level curriculum co-training strategy. *Condition-level curriculum.* We introduce conditioning modalities in three stages of increasing difficulty. We begin with text-conditioning, the least difficult task, as it closely aligns with the base model’s generation ability. Subsequently, we progressively incorporate the more challenging modalities: first, reference-video conditioning, and finally, the most demanding task, trajectory conditioning. This staged approach allows the model to gradually adapt from familiar semantic guidance to complex geometric control. *Data-level curriculum.* We devise a two-substage training process to combine the strengths of diverse data sources. First, extensive training on large-scale synthetic data (UE5) establishes precise camera control. This is followed by a brief fine-tuning on a small set of real videos to restore photorealism. This short adaptation rapidly leverages the model’s inherent visual priors, correcting the realism gap from synthetic data without causing catastrophic forgetting of the learned motion control. Furthermore, we alleviate modality conflicts using Condition RoPE, which explicitly encodes and separates conditioning inputs.

In summary, our contributions are:

- We propose **OmniCamera**, the first unified video generation model that seamlessly integrates three camera conditions (text, trajectory, reference video) with three content conditions (text, image, video). This framework enables arbitrary condition combinations with high parameter efficiency.
- We curate the **OmniCAM** dataset, a hybrid dataset combining high-precision synthetic camera trajectories with diverse real-world videos, providing robust multi-level supervision for generalized camera control.
- We introduce a **Dual-level Curriculum Co-Training** strategy, comprising a Condition-level Curriculum to avoid modality conflicts and a Data-level

Table 1: Method comparison. OmniCamera decouples video generation into independent content conditions and camera conditions, enabling arbitrary combinations.

Method	Content			Camera		
	Text	Image	Video	Text	Traj.	Video
CameraCtrl [17]	✓	✓	×	×	✓	×
AC3D [1]	✓	×	×	×	✓	×
RealCam [22]	×	✓	×	×	✓	×
FloVD [20]	×	✓	×	×	✓	×
CamCloneMaster [26]	×	✓	✓	×	×	✓
ReCamMaster [2]	×	×	✓	×	✓	×
UNIC [36]	×	×	✓	×	✓	×
OmniVCus [8]	×	×	✓	×	✓	×
Wan-Fun-Camera [31]	×	✓	×	✓	×	×
Ours	✓	✓	✓	✓	✓	✓

Table 2: Comparison of camera-controllable datasets. OmniCAM is the only dataset providing comprehensive data types for both synthetic and real-world scenarios. **SS**: Same-Scene, Diverse-Trajectory; **ST**: Same-Trajectory, Diverse-Scene.

Dataset	Data Types					Preference	
	Text	Traj.	SS-Data	ST-Data	Triplet	#Videos	Resolution
<i>Syn.</i>	SynCamMaster [3]	×	✓	✓	×	×	34K 1280×1280
	ReCamMaster [2]	×	✓	✓	×	×	136K 1280×1280
	CamCloneMaster [26]	×	×	×	✓	✓	391K 1008×576
	Ours (UE5)	✓	✓	✓	✓	✓	500K 1280×1280
	RealEstate-10K [40]	×	✓	×	×	×	10K 1280×720
<i>Real</i>	DL3DV-10K [24]	×	✓	×	×	×	10.5K 3840×2160
	ScanNet [10]	×	✓	×	×	×	1513 1296×968
	Matterport3D [9]	×	✓	×	×	×	90 1280×1024
	ARKitScenes [4]	×	✓	×	×	×	1661 1920×1440
	Ours (Real)	✓	✓	✓	✓	✓	330K 3840×2160

Curriculum to bridge the domain gap between synthetic and real data, ensuring stable and effective multimodal learning.

Comprehensive quantitative and qualitative experiments demonstrate that our approach surpasses models trained under single-task or single-condition settings, highlighting the effectiveness of the proposed unified framework for camera-controlled video generation.

2 Related Works

Video Generation. Recent advances in video generation have led to rapid progress, with numerous studies focusing on text-to-video (T2V) [6, 7, 13, 18, 27, 31] and image-to-video (I2V) [13, 31, 32, 35] generation. Inspired by the success of powerful image generation models such as Stable Diffusion [11] and Flux [21], recent methods have widely adopted Diffusion Transformers (DiT) [28] and flow matching [25] architectures for video generation. Modern approaches increasingly aim to unify text and image conditioning within a single framework. For example, Wan 2.2 [31] introduces a TI2V model that replaces the first-frame latent with image embeddings for I2V tasks, while Seedance [13] jointly trains T2V and I2V models and integrates video-specific RLHF with multi-dimensional reward mechanisms to enhance quality.

Camera-controlled Video Generation. Existing methods incorporate various camera motion control conditions into T2V and I2V models, including textual descriptions [15, 31], 3D trajectories [1, 1, 16, 17, 20, 22], and reference videos [26]. CameraCtrl [16] trains a camera adaptor integrated with T2V and I2V models to inject trajectory-based camera motion, while AC3D [1] explores camera motion representations within diffusion transformers to achieve controllable generation. Several studies [12, 19, 20, 22, 29, 33, 38] further leverage expert models such as depth [29, 34] and optical flow [20] networks to provide geometric priors via point clouds or motion fields. However, the effectiveness of motion control remains bounded by the precision of expert models and the overhead of computing explicit geometric information.

Camera-controlled Video-to-Video Generation. Camera-controllable V2V generation aims to synthesize a new video by combining the visual content of

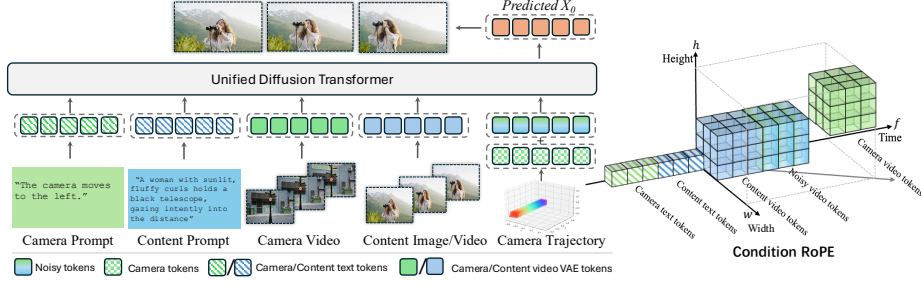


Fig. 3: Pipeline of OmniCamera. Left: Diffusion Transformer with *decoupled* condition injection: text and reference video tokens are concatenated for self-attention; trajectory features are added to the noise latent. Right: Condition RoPE explicitly disentangles modalities using unique positional encodings.

a given video with a novel camera motion hint [2, 5, 14, 26, 30, 37, 39]. Early studies explored this task using Kubric-simulated data [30] or by leveraging additional geometric cues such as 3D point tracking [5, 37]. Recent works, such as ReCamMaster [2], directly take a content video and a new trajectory as inputs to re-generate videos by training T2V models on carefully curated datasets. CamCloneMaster [26] further extends this idea by introducing a video re-shot model capable of guiding both I2V and V2V generation with video hints.

3 Method

Fig. 3 provides an overview of our OmniCamera framework. We begin by outlining the preliminaries in Sec. 3.1. Our method is then presented through four key components: *Dataset Design* (Sec. 3.2), *Model Architecture* (Sec. 3.3), *Multi-condition Co-Training* (Sec. 3.4), and *Multi-task Inference* (Sec. 3.5).

3.1 Preliminary

We first introduce the fundamental concept of Flow Matching [25], which has been widely adopted in recent video generation models [31]. Flow Matching formulates generative modeling as learning a continuous velocity field that transports a simple prior distribution (e.g., Gaussian) to the target data distribution. Given a data sample $x_0 \sim p_{\text{data}}$ and a noise sample $x_1 \sim p_{\text{prior}}$, Flow Matching constructs a linear interpolant:

$$x_t = (1 - t)x_0 + tx_1, \quad t \in [0, 1]. \quad (1)$$

The corresponding target velocity field is defined as:

$$u_t(x_t) = x_1 - x_0, \quad (2)$$

which is independent of timestep t but conditioned on the pair (x_0, x_1) . The training objective is to optimize a neural network $v_\theta(x_t, t)$ to approximate this conditional velocity field, thereby recovering the probability flow between the prior and data distributions.

3.2 Dataset Design

Training our unified model requires diverse supervision signals: single videos with trajectories or textual motion annotations, same-scene video pairs with different camera motions, cross-scene pairs with matched motions, and video triplets. We construct the **OmniCAM** dataset, a hybrid dataset from synthetic and real-world sources, as illustrated in Fig. 4.

Synthetic Data. We define a library of approximately 50 camera-motion types (20 basic and 30 complex), each associated with a detailed textual description, and synthesize corresponding videos in UE5 following [2, 26]. As shown in the left of Fig. 4, the synthetic dataset is categorized into three distinct subsets:

(1) **Same-Scene, Diverse-Trajectory Data:** Within a fixed *Scene 1*, we first establish a set of camera movement rules to automatically batch-generate diverse camera trajectories. We then simulate a simultaneous shooting process by positioning multiple cameras that face the subjects and move along these pre-defined trajectories. This allows us to render datasets with synchronized cameras capturing objects from varying perspectives.

(2) **Same-Trajectory, Diverse-Scene Data:** We place random subjects performing arbitrary actions in *Scene 1* and *Scene 2*. We then render these scenarios by applying identical camera trajectories across these scenes, ensuring consistent camera motion patterns.

(3) **Motion-Content-Target Triplets:** We construct training triplets denoted by $\{z_m, z_d, z_0\}$, where z_0 represents the target video, which shares identical camera motion with the motion video z_m (derived from the same-trajectory, diverse-scene subset) and shares identical content with the content video z_d (derived from the same-scene, diverse-trajectory subset).

In total, we sample 250K distinct camera trajectories (5K trajectories for each of the 50 predefined motion categories) to render 500K independent video clips with accurate pose annotations. By combining these clips, we construct 500K same-scene diverse-trajectory pairs, 500K same-trajectory diverse-scene pairs, and 500K motion-content-target triplets, thereby providing massive and diverse paired supervision.

Real-World Data. In addition to synthetic data, we curate a real-world dataset through a rigorous pipeline. As illustrated in the right of Fig. 4, our data processing pipeline comprises four key steps:

(1) **Trajectory Estimation:** We utilize MegaSaM [23] to extract camera trajectories, followed by camera parameter calibration from CameraCtrl2 [17].

(2) **Trajectory Filtering:** Since raw trajectories extracted by MegaSaM often contain significant noise, we filter them based on trajectory smoothness. Let $\mathbf{c}_i$ denote the camera position at frame i (where $i \in \{1, \dots, N\}$ for a video of N frames) and $d_i = \|\mathbf{c}_{i+1} - \mathbf{c}_i\|_2$ the frame-to-frame displacement. We compute:

$$r_{\text{jump}} = \frac{\max_i d_i}{\bar{d}}, \quad r_{\text{comp}} = \frac{L}{\|\mathbf{c}_N - \mathbf{c}_1\|_2 + \epsilon}, \quad (3)$$

where $\bar{d} = \frac{1}{N-1} \sum_{i=1}^{N-1} d_i$ and $L = \sum_{i=1}^{N-1} d_i$. We apply two criteria: (i) *jump filtering*, discarding trajectories with $r_{\text{jump}} > \tau_{\text{jump}}$; and (ii) *complexity filtering*,

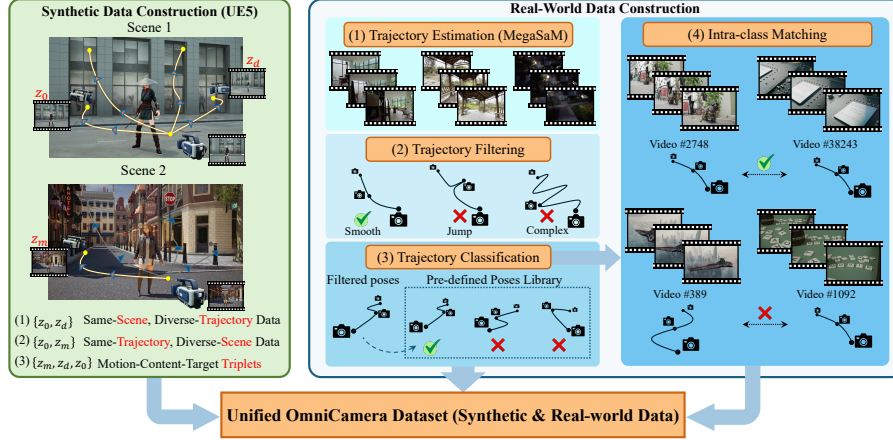


Fig. 4: OmniCAM dataset construction. **Left:** UE5 synthetic videos provide accurate camera poses to build paired and triplet supervision. **Right:** real videos are processed via trajectory estimation, trajectory filtering, trajectory classification, and intra-class matching to obtain reliable trajectories and cross-scene motion pairs.

discarding trajectories with $r_{\text{comp}} > \tau_{\text{comp}}$, which indicates overly tortuous or jittery motion. Here, τ_{jump} and τ_{comp} are preset thresholds, and ϵ is a small constant for numerical stability.

(3) Trajectory Classification: We categorize the filtered trajectories into 50 predefined camera motion patterns based on similarity. For each video, we compute the trajectory similarity with all 50 predefined templates based on translation error (TransErr) and rotation error (RotErr) [16], and assign it to the class yielding the minimum error.

(4) Intra-class Matching: Within each class, we perform random pairwise matching of trajectories. We compute the TransErr and RotErr between two candidate trajectories and consider them as a valid match only if both errors are below strict predefined thresholds. This process yields real-world data pairs characterized by identical trajectories but diverse scenes.

Through this progressive pipeline, we obtain data at three levels of annotation: 380K videos with reliable trajectory annotations from Step (2), a total of 380K videos augmented with categorical motion descriptions from Step (3), and 300K Same-Trajectory, Diverse-Scene real-video pairs identified from Step (4).

3.3 Model Design

To avoid modality conflicts, we adopt a decoupled condition-injection strategy, as shown in the left of Fig. 3: textual prompts and visual conditions (content latent z_d and camera-motion latent z_m) are unified as sequence-level representations to interact with the noise latent z_t during joint attention; concurrently, trajectory parameters are processed via an MLP before being passed into DiT blocks.

3D Condition RoPE. As illustrated in the right of Fig. 3, to resolve the spatial-temporal ambiguity caused by sequence concatenation, we propose a 3D

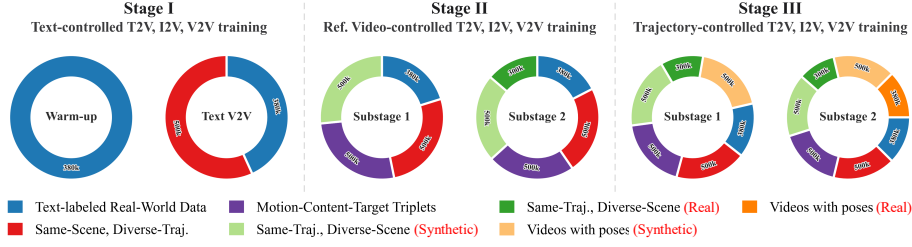


Fig. 5: Dual-level curriculum data composition. Condition-level curriculum progresses from text control (Stage I) to reference-video control (Stage II) to trajectory control (Stage III), following a coarse-to-fine and easy-to-hard order. In Stage II and Stage III, data-level curriculum first trains on synthetic data with precise trajectories (Substage 1), then fine-tunes on real-world videos to recover photorealism (Substage 2).

Condition RoPE. For an original token coordinate (f, h, w) denoting the frame, height, and width index, we assign distinct spatial-temporal base offsets for each modality: $(0, 0, 0)$ for z_t , $(F, 0, 0)$ for z_d , and (F, H, W) for z_m , where F, H, W represent the total frames, height, and width dimensions of the noise latent z_t . The positional codes are then computed on these shifted coordinates, with frequencies $\text{freq}_i = \theta^{-2i/D}$, where D is the channel dimension of the token embeddings and θ is the base constant. This explicitly distinguishes identical coordinate structures, seamlessly integrating multiple conditions within our unified framework.

3.4 Dual-level Curriculum Co-Training

Simultaneously training 9 different generation combinations (3 camera-control modalities $\times$ 3 generative tasks) introduces severe representation conflicts and optimization instability. Naively mixing all data and conditions from scratch leads to degraded camera controllability and deteriorated visual quality. To address this, we propose a **Dual-level Curriculum Co-Training** strategy operating along two complementary axes: a task-level curriculum to align multi-modal conditions progressively, and a data-level curriculum to balance camera controllability and visual realism.

Condition-level Curriculum. As shown in Fig.5, we design a three-stage curriculum that evolves from coarse-to-fine granularity and easy-to-hard difficulty. In **Stage I**, we focus on text-based conditioning (across T2V, I2V, and V2V tasks). This acts as semantic guidance with the coarsest control granularity. Since it aligns closely with the base model’s inherent capabilities, this stage serves as the foundational and easiest learning step. In **Stage II**, we incorporate reference-video control. Operating at a relatively coarse granularity, this stage employs in-context learning to transfer coarse global camera motion from a reference video to the target. This introduces a higher level of learning difficulty compared to pure text guidance. In **Stage III**, we introduce camera trajectory control, which demands the finest granularity and precise geometric control. Consequently, this is the most challenging modality. By adhering to this coarse-to-fine and easy-to-hard progression, we effectively prevent optimization collapse and achieve reliable multi-condition control.

Data-level Curriculum. High-quality multi-condition training faces a severe data dilemma: simulated data (e.g., UE5) provides perfectly accurate trajectories but suffers from a significant domain gap compared to real-world videos; conversely, real-world data possesses high visual realism but yields noisy trajectory estimates via Structure-from-Motion (SfM). To integrate the advantages of diverse data sources, we develop a two-substage data curriculum applied in the later conditioning stages (Stage II and Stage III). In **Substage 1**, we utilize a large scale of UE5 data paired with precise trajectories to significantly boost camera motion accuracy. In **Substage 2**, we employ a curated set of high-quality real videos with reliable trajectory estimates to restore the model’s capability in generating photorealistic content.

3.5 Multi-task Inference

We employ a dual-condition classifier-free guidance strategy. The model relies on text C_T for semantic control and multimodal inputs C_M for camera motion. To enable flexible controllability, we randomly drop conditions during training with a probability of 5% for C_M only, 5% for C_T only, and 5% for both simultaneously. At the inference stage, this allows us to utilize separate guidance scales, w_T and w_M , to independently adjust the influence of text and motion guidance. The modified score estimate is as follows:

$$\begin{aligned}\hat{\epsilon}_\theta(z_t, c_T, c_M) &= \epsilon_\theta(z_t, \phi, \phi) \\ &\quad + w_T (\epsilon_\theta(z_t, c_T, \phi) - \epsilon_\theta(z_t, \phi, \phi)) \\ &\quad + w_M (\epsilon_\theta(z_t, c_T, c_M) - \epsilon_\theta(z_t, c_T, \phi))\end{aligned}\tag{4}$$

Here, $c_M \in \{\text{Trajectory, Ref-Video}\}$ acts as a polymorphic camera condition.

4 Experiments

4.1 Implementation Details

We build upon the 5B-parameter Wan2.2-TI2V architecture, utilizing its VAE for 704×1248 video generation. During training, we exclusively optimize the self-attention, cross-attention, and camera embedding modules. We train the model on 32 H100 GPUs for 60K steps with a batch size of 32, using the AdamW optimizer (initial learning rate 7×10^{-5} decaying to 5×10^{-5} via a cosine schedule). **Evaluation.** We have constructed a comprehensive evaluation dataset that comprises 36 camera-motion text commands, 36 camera trajectories, and 36 reference videos. For each task, we assess the camera-control performance on 800 generated videos. The evaluation metrics include CLIP-T, CLIP-F, Rotation Error, Translation Error, FVD or FVD-V, as well as motion accuracy. Detailed calculations for these metrics can be found in the Supplementary Material.

4.2 Text Condition

As shown in Fig. 6, text-controlled generation provides semantic-level camera guidance by prepending motion instructions to the input caption. As shown



Fig. 6: Qualitative results for **text-controlled** camera motion.

Table 3: Quantitative results of **Text-controlled** camera motion on T2V, I2V and V2V tasks. **Table 4:** Quantitative results of **Trajectory-controlled** camera motion on the tasks (CLIP-T for T2V/I2V, CLIP-V for V2V).

Task	Method	FVD↓	CLIP-F↑	CLIP-T↑	MotionAcc↑
T2V	AnimateDiff [15]	1471.28	96.39	25.04	19.3%
	Wan2.1-T2V [31]	899.21	98.54	27.75	28.9%
	Wan2.2-TI2V [31]	867.27	98.92	28.06	31.4%
	Ours	884.11	98.94	29.06	92.4%
I2V	Wan2.1-I2V [31]	345.69	98.27	29.95	27.5%
	Wan2.2-TI2V [31]	285.52	99.05	30.30	30.3%
	Wan2.2-Fun-Camera	314.63	99.02	30.08	68.8%
	Ours	308.83	99.05	30.29	91.3%
V2V	Ours	430.71	99.02	30.17	89.7%

Task	Method	FVD↓	RotErr↓	TransErr↓	CLIP(T/V)↑
T2V	CameraCtrl [16]	1365.90	0.126	8.160	23.43
	AC3D [1]	958.62	0.056	7.542	30.01
	Ours	893.95	0.034	2.064	29.64
I2V	CameraCtrl [16]	586.14	0.048	3.798	27.04
	Ours	333.54	0.043	3.355	30.40
V2V	ReCamMaster [2]	366.82	0.048	5.320	94.10
	Ours	331.36	0.045	5.933	94.97

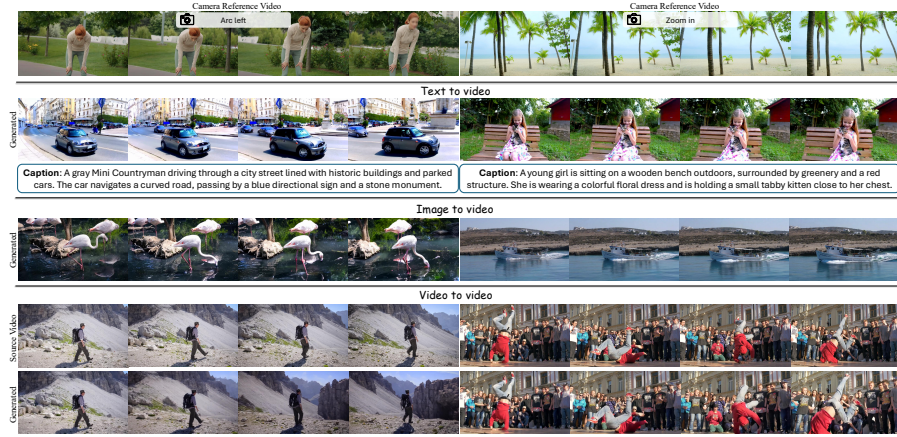
in Tab. 3, foundational video models (e.g., AnimateDiff [15] and Wan2.2 [31]) struggle to execute specific camera commands. Even the specialized Wan2.2-Fun-Camera model can only perform simple camera movements (e.g., pan and tilt) with suboptimal precision. As further visualized in Fig. 9, Wan2.2-Fun-Camera also suffers from severe content hallucination and structural artifacts (e.g., semantic confusion in animals and static volcanic smoke). In comparison, OmniCamera supports complex 3D spatial movements and outperforms all baselines in motion accuracy by a large margin while maintaining competitive visual fidelity. Additionally, OmniCamera uniquely enables *text-guided* camera motion editing on existing videos (V2V), a capability under-explored in prior models.

4.3 Reference-Video Condition

As shown in Fig. 7 and Tab. 5, OmniCamera achieves accurate reference-video-controlled camera motion across T2V, I2V, and V2V tasks. Compared with CamCloneMaster [26] (Fig. 9), OmniCamera yields significantly more accurate camera motion, superior content preservation, and fewer visual artifacts. This

Table 5: Quantitative results of **Reference-video-controlled** camera motion on T2V, I2V, and V2V tasks. “-” means methods fail to accomplish task.

Methods	T2V				I2V				V2V			
	FVD↓	RotErr↓	TransErr↓	CLIP↑	FVD↓	RotErr↓	TransErr↓	CLIP↑	FVD-V↓	RotErr↓	TransErr↓	CLIP-V↑
CamCloneMaster [26]	-	-	-	-	380.73	0.021	3.953	30.18	376.33	0.019	4.936	93.60
Ours	868.63	0.024	7.109	29.30	352.76	0.023	4.195	30.45	348.51	0.019	5.392	93.71

**Fig. 7:** Qualitative results for **reference-video-controlled** camera motion.

strong performance is directly attributed to the proposed dual-level curriculum training strategy and our superior data composition, which combines precise UE5 synthetic data with diverse real-world video pairs.

4.4 Trajectory Condition

As shown in Fig. 8 and Tab. 4, OmniCamera achieves highly accurate 3D trajectory control across T2V, I2V, and V2V tasks. Compared with state-of-the-art methods such as ReCamMaster [2] (Fig. 9), our approach exhibits superior motion precision with fewer visual artifacts. Benefiting from our OmniCAM dataset and dual-level curriculum co-training, OmniCamera closely follows diverse trajectories and generalizes to complex paths (e.g., triangular and polyline paths).

4.5 Combining Control Modalities and Conflict Analysis

Multiple control modalities can be combined to achieve composite camera effects. As shown in Fig. 2, distinct conditions can be seamlessly integrated; for instance, a semantic text command (e.g., “Pan right”) can be combined with a 3D trajectory (e.g., dolly out), producing a fused motion in the generated video. We further test potential modality conflicts by providing contradictory motions in pairwise combinations of trajectory, reference video, and text conditions. As shown in Fig. 10, we observe a consistent dominance hierarchy: **Trajectory** > **Reference Video** > **Text**, indicating that conditions with stronger geometric

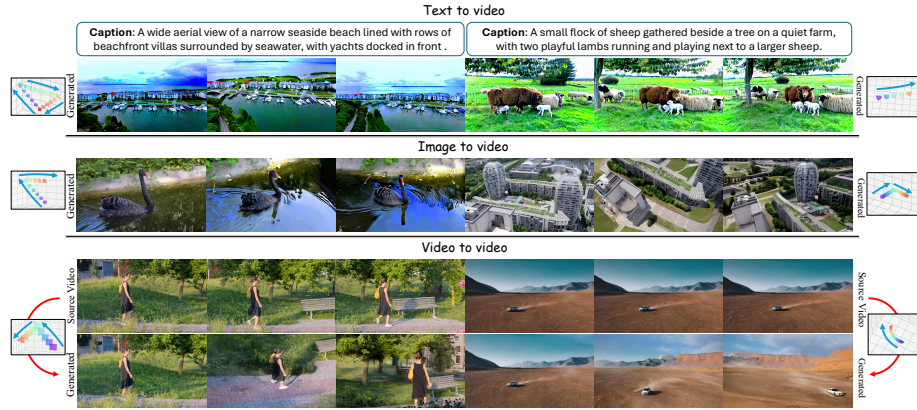


Fig. 8: Qualitative results for **trajectory-controlled** camera motion.

Table 6: **Ablation study** of Curriculum Co-Training across different tasks. **CC**: condition-level curriculum, **DC**: data-level curriculum.

Method	Trajectory-controlled									Reference-video-controlled								
	T2V			I2V			V2V			T2V			I2V			V2V		
	TransErr↓	RotErr↓	FVD↓	TransErr↓	RotErr↓	FVD↓	TransErr↓	RotErr↓	FVD-V↓	TransErr↓	RotErr↓	FVD↓	TransErr↓	RotErr↓	FVD↓	TransErr↓	RotErr↓	FVD-V↓
w/o CC	3.512	0.052	1024.31	4.821	0.061	412.50	7.425	0.068	405.12	9.510	0.038	982.45	5.762	0.035	410.28	7.214	0.029	415.82
w/o DC	2.845	0.045	965.84	4.103	0.052	378.42	6.512	0.055	362.75	8.324	0.031	925.61	4.951	0.028	381.54	6.185	0.024	382.46
Ours	2.064	0.034	893.95	3.355	0.043	333.54	5.933	0.045	331.36	7.109	0.024	868.63	4.195	0.023	352.76	5.392	0.019	348.51

Table 7: **Ablation study** of Data Composition across different tasks. **UE5**: UE5 synthetic data, **Real**: real-world data.

Method	Trajectory-controlled									Reference-video-controlled								
	T2V			I2V			T2V			I2V			T2V			I2V		
	TransErr↓	RotErr↓	FVD↓	TransErr↓	RotErr↓	FVD↓	TransErr↓	RotErr↓	FVD↓	TransErr↓	RotErr↓	FVD↓	TransErr↓	RotErr↓	FVD↓	TransErr↓	RotErr↓	FVD↓
w/o UE5	3.105	0.050	915.22	4.512	0.058	350.14	8.845	0.035	890.56	5.321	0.030	365.21	7.315	0.026	1085.24	4.352	0.025	492.35
w/o Real	2.152	0.036	1150.45	3.481	0.045	485.62	7.315	0.026	1085.24	4.352	0.025	492.35	7.315	0.026	1085.24	4.352	0.025	492.35
Ours	2.064	0.034	893.95	3.355	0.043	333.54	7.109	0.024	868.63	4.195	0.023	352.76	7.109	0.024	868.63	4.195	0.023	352.76

explicitness naturally override weaker, semantic-level ones. For instance, when the input condition combines a “pan left” trajectory with a “pan right” reference video, the output strictly follows the “pan left” camera motion dictated by the trajectory. Similarly, when combining a “pan right” reference video with a “pan left” text prompt, the output adheres exclusively to the “pan right” motion dictated by the video.

4.6 Ablation Study

We perform ablation studies to analyze our dual-level curriculum, including the *condition-level curriculum* (**CC**) and the *data-level curriculum* (**DC**), as well as the contribution of OmniCAM’s hybrid data sources by ablating either UE5 synthetic videos or curated real-world videos.

Effect of Dual-level Curriculum. Tab. 6 validates the necessity of our dual-level curriculum. *w/o CC* jointly trains all tasks from scratch; this naive mixing causes representation conflicts that fail to disentangle camera and content conditions. According to the qualitative results in Fig. 11(b)(c), removing CC leads to



Fig. 9: Visual comparison of OmniCamera with state-of-the-art methods. **Text Control:** Wan2.2-Fun-Camera produces some artifacts (e.g., the “ear” mutating into “eye”). **Trajectory Control:** ReCamMaster [2] produces severe distortions on the “bus” and “motorcyclist”. **Reference-Video Control:** CamCloneMaster [26] executes incorrect camera motion while introducing heavy background and object distortions.

severe visual degradation, including frame deterioration (e.g., in the car scene) and content hallucination (e.g., erroneous trees). Furthermore, according to the qualitative results *w/o DC* in Fig. 11(a), it hinders the learning of precise pose control, e.g., missing the “pan left” in the camel scene and “boom down, tilt up” in the bear scene, whereas ours strictly follows the desired trajectories.

Effect of Data Composition. Tab. 7 explores the effect of data sources by removing UE5 data (*w/o UE5*) or real data (*w/o Real*). We observe that UE5 data mainly contributes to camera controllability, whereas real data is crucial for closing the domain gap and recovering photorealism. According to the qualitative results of removing real data in Fig. 11(d), relying solely on UE5 data (*w/o Real*) results in a noticeable drop in visual realism and introduces erroneous artifacts such as grids in red boxes. In contrast, training with both data sources seamlessly preserves the learned camera control while yielding high-fidelity appearances.

Further ablations. Please refer to the supplementary materials for additional ablations on 3D Condition RoPE and dual-condition CFG.



Fig. 10: Modality conflict analysis. Pairwise combination of contradictory camera conditions reveals a dominance order: trajectory>reference video>text prompts.

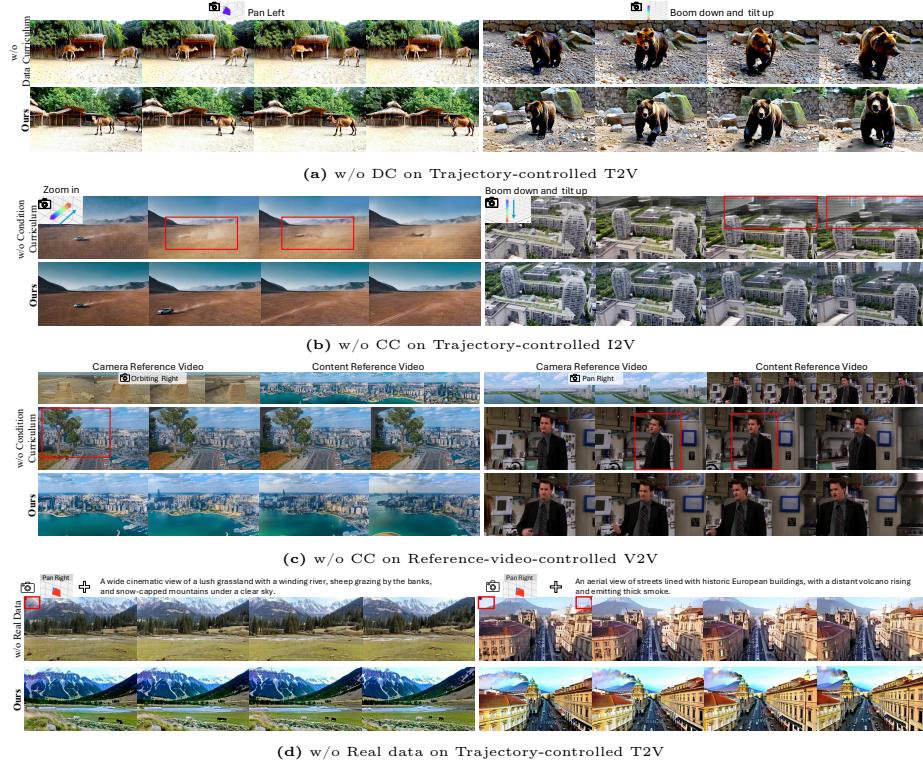


Fig. 11: Qualitative ablation. (a) Removing DC leads to failed camera motions and degraded visual quality. (b)(c) Omitting CC results in structural distortions in I2V and content hallucinations in V2V (red boxes). (d) Training solely on synthetic data ensures precise control but lacks photorealism.

5 Conclusion

This paper presents **OmniCamera**, a unified multi-task video generation framework that supports text, trajectory, and reference-video camera control across T2V, I2V, and V2V. We construct the **OmniCAM** dataset that combines high-precision synthetic trajectories with real-world videos, and propose a dual-level curriculum co-training strategy for stable multi-condition learning. We further introduce 3D Condition RoPE and a dual-condition CFG to mitigate condition conflicts and improve camera controllability. Extensive experiments demonstrate improved controllability with competitive visual quality across tasks.

Limitation. OmniCamera successfully explores and demonstrates the feasibility of decoupling the observation perspective (camera) and scene content in video generation. However, while we unify the most common representations for these two dimensions, our current framework does not yet accommodate finer-grained controls, such as multiple reference images or localized motion guidance. We leave these extensions for future work.

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 (2025)
2. 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 (2025)
3. Bai, J., Xia, M., Wang, X., Yuan, Z., Fu, X., Liu, Z., Hu, H., Wan, P., Zhang, D.: Syncammaster: Synchronizing multi-camera video generation from diverse view-points. arXiv preprint arXiv:2412.07760 (2024)
4. Baruch, G., Chen, Z., Dehghan, A., Dimry, T., Feigin, Y., Fu, P., Gebauer, T., Joffe, B., Kurz, D., Schwartz, A., Shulman, E.: ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1) (2021), https://openreview.net/forum?id=tjZjv_qh_CE
5. Bian, W., Huang, Z., Shi, X., Li, Y., Wang, F.Y., Li, H.: Gs-dit: Advancing video generation with pseudo 4d gaussian fields through efficient dense 3d point tracking. arXiv preprint arXiv:2501.02690 (2025)
6. 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)
7. 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)
8. Cai, Y., Zhang, H., Chen, X., Xing, J., Hu, Y., Zhou, Y., Zhang, K., Zhang, Z., Kim, S.Y., Wang, T., et al.: Omnivcus: Feedforward subject-driven video customization with multimodal control conditions. arXiv preprint arXiv:2506.23361 (2025)
9. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. International Conference on 3D Vision (3DV) (2017)
10. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR (2017)
11. 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)
12. Feng, W., Liu, J., Tu, P., Qi, T., Sun, M., Ma, T., Zhao, S., Zhou, S., He, Q.: I2vcontrol-camera: Precise video camera control with adjustable motion strength. arXiv preprint arXiv:2411.06525 (2024)

13. Gao, Y., Guo, H., Hoang, T., Huang, W., Jiang, L., Kong, F., Li, H., Li, J., Li, L., Li, X., et al.: Seedance 1.0: Exploring the boundaries of video generation models. arXiv preprint arXiv:2506.09113 (2025)
14. Gu, Z., Yan, R., Lu, J., Li, P., Dou, Z., Si, C., Dong, Z., Liu, Q., Lin, C., Liu, Z., et al.: Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (2025)
15. 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)
16. 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)
17. 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)
18. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868 (2022)
19. Hou, C., Chen, Z.: Training-free camera control for video generation. arXiv preprint arXiv:2406.10126 (2024)
20. Jin, W., Dai, Q., Luo, C., Baek, S.H., Cho, S.: Flovd: Optical flow meets video diffusion model for enhanced camera-controlled video synthesis. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2025)
21. Labs, B.F.: Flux. <https://github.com/black-forest-labs/flux> (2024)
22. Li, T., Zheng, G., Jiang, R., Zhan, S., Wu, T., Lu, Y., Lin, Y., Deng, C., Xiong, Y., Chen, M., et al.: Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. arXiv preprint arXiv:2502.10059 (2025)
23. Li, Z., Tucker, R., Cole, F., Wang, Q., Jin, L., Ye, V., Kanazawa, A., Holynski, A., Snavely, N.: MegaSaM: Accurate, fast and robust structure and motion from casual dynamic videos. In: CVPR (2025)
24. Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., et al.: D13dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In: CVPR (2024)
25. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
26. Luo, Y., Bai, J., Shi, X., Xia, M., Wang, X., Wan, P., Zhang, D., Gai, K., Xue, T.: Camclonemaster: Enabling reference-based camera control for video generation. arXiv preprint arXiv:2506.03140 (2025)
27. Ma, X., Wang, Y., Chen, X., Jia, G., Liu, Z., Li, Y.F., Chen, C., Qiao, Y.: Latte: Latent diffusion transformer for video generation. Transactions on Machine Learning Research (2025)
28. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
29. 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)
30. 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)
31. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
 32. Xing, J., Xia, M., Zhang, Y., Chen, H., Yu, W., Liu, H., Liu, G., Wang, X., Shan, Y., Wong, T.T.: Dynamicrafter: Animating open-domain images with video diffusion priors. In: European Conference on Computer Vision. pp. 399–417. Springer (2024)
 33. 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)
 34. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10371–10381 (2024)
 35. 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)
 36. Ye, Z., He, X., Liu, Q., Wang, Q., Wang, X., Wan, P., Zhang, D., Gai, K., Chen, Q., Luo, W.: Unic: Unified in-context video editing. arXiv preprint arXiv:2506.04216 (2025)
 37. YU, M., Hu, W., Xing, J., Shan, Y.: Trajectoryrafter: Redirecting camera trajectory for monocular videos via diffusion models. arXiv preprint arXiv:2503.05638 (2025)
 38. Yu, W., Xing, J., Yuan, L., Hu, W., Li, X., Huang, Z., Gao, X., Wong, T.T., Shan, Y., Tian, Y.: Viewrafter: Taming video diffusion models for high-fidelity novel view synthesis. arXiv preprint arXiv:2409.02048 (2024)
 39. Zhang, D.J., Paiss, R., Zada, S., Karnad, N., Jacobs, D.E., Pritch, Y., Mosseri, I., Shou, M.Z., Wadhwa, N., Ruiz, N.: Recapture: Generative video camera controls for user-provided videos using masked video fine-tuning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2050–2062 (2025)
 40. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. arXiv preprint arXiv:1805.09817 (2018)