RealCam-I2V: Real-World Image-to-Video Generation with Interactive Complex Camera Control

Teng Li[1*], Guangcong Zheng[1*], Rui Jiang[1], Shuigen Zhan[1], Tao Wu[1], Yehao Lu[1], Yining Lin[3], Chuanyun Deng[2], Yepan Xiong[2], Min Chen[2], Lin Cheng[2], and Xi Li[1†]

[1] College of Computer Science & Technology, Zhejiang University
[2] Central Media Technology Institute, 2012 Lab, Huawei
[3] Supremind

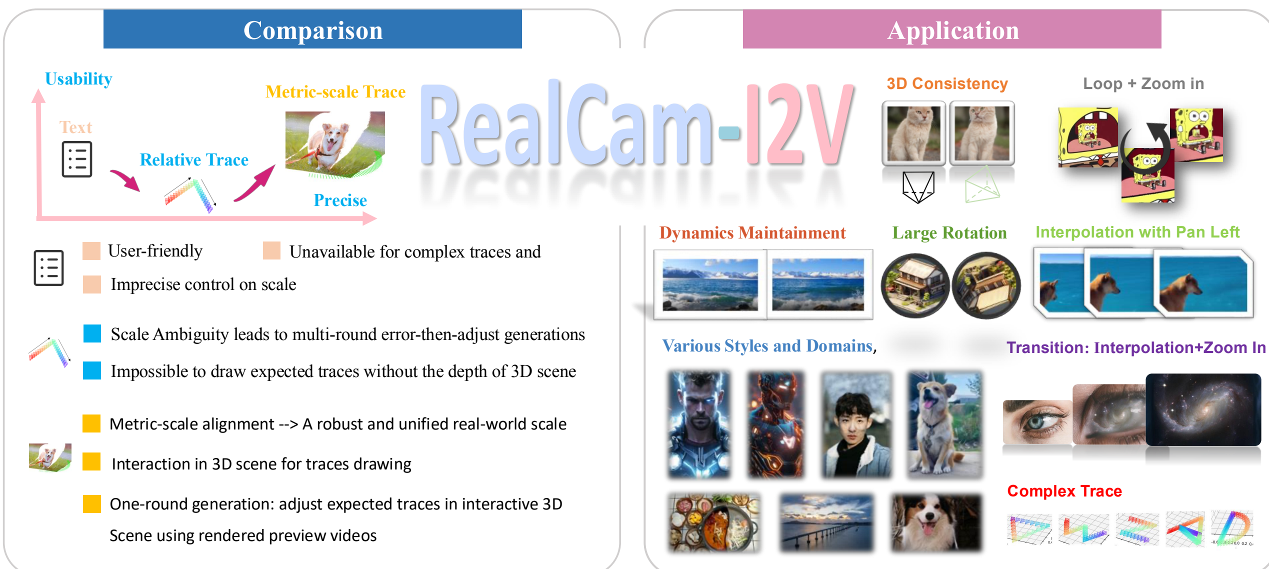

Figure 1. We propose RealCam-I2V, a camera controllable image-to-video generation framework for complex real-world camera control and extra applications including camera-controlled loop video generation, generative frame interpolation, and smooth scene transitions.

Abstract

Recent advancements in camera-trajectory-guided image-to-video generation offer higher precision and better support for complex camera control compared to text-based approaches. However, they also introduce significant usability challenges, as users often struggle to provide precise camera parameters when working with arbitrary real-world images without knowledge of their depth nor scene scale. To address these real-world application issues, we propose RealCam-I2V, a novel diffusion-based video generation framework that integrates monocular metric depth estimation to establish 3D scene reconstruction in a preprocessing step. During training, the reconstructed 3D scene enables scaling camera parameters from relative to metric scales, ensuring compatibility and scale consistency across diverse real-world images. In inference, RealCam-I2V offers an intuitive interface where users can precisely draw camera trajectories by dragging within the 3D scene. To further enhance precise camera control and scene consistency, we propose scene-constrained noise shaping, which shapes high-level noise and also allows the framework to maintain dynamic and coherent video generation in lower noise stages. RealCam-I2V achieves significant improvements in controllability and video quality on the RealEstate10K and out-of-domain images. We further enables applications like camera-controlled looping video generation and generative frame interpolation. Project page: `zgctroy.github.io/RealCam-I2V`.

*Equal Contribution

1. Introduction

Recent advancements in image-to-video generation [9, 10, 20, 24, 82] have significantly improved controllability over synthesized videos. However, challenges remain in achieving realistic, controllable camera movement within complex real-world scenes. Text-based camera-control methods [4, 20, 29, 31, 38, 66], like traditional diffusion-based video generation, are intuitive and straightforward but lack precision in explicit control over camera parameters, such as angle, scale, and movement direction. This limitation has spurred the development of camera-trajectory-guided approaches, which attempt to address these issues by offering finer control over camera movement.

Current camera-trajectory-guided methods typically rely on relative camera trajectories, as seen in models like MotionCtrl [69], CameraCtrl [22], CamCo [78], and CamI2V [96]. While these approaches provide more control than text-based models, they are fundamentally limited by their reliance on relative scale trajectories. Training on relative scales results in inconsistencies when applied to real-world scenes, where metric scale is crucial for realistic depth perception. Additionally, without access to depth information, users find it challenging to draw precise trajectories, making these methods difficult to use effectively.

To overcome these limitations, we propose RealCam-I2V, a video generation framework that integrates monocular depth estimation as a preprocessing step to construct a robust, metric-scale 3D scene. Our approach leverages the Depth Anything v2 [80] model (metric version) to predict the metric depth of a user-provided reference image, reprojecting its pixels back into camera space to create a stable 3D representation. This 3D scene serves as the foundation for camera control, providing a consistent and metric scale that is critical for real-world applications.

In the training stage, we align the reconstructed 3D scene of the reference image with the point cloud of each video sample, reconstructed using COLMAP [54], a structure-from-motion (SfM) method. This alignment allows us to rescale COLMAP-annotated camera parameters to the Depth Anything metric, providing an metric, stable, and robust scale across training data. By aligning relative-scale camera parameters to metric scales, we can condition the video generation model on accurately scaled camera trajectories, achieving greater control and scene consistency across diverse real-world images.

During inference, RealCam-I2V provides an interactive interface where users can intuitively design camera trajectories by drawing within the reconstructed 3D scene of the reference image. This interface renders preview videos of the trajectory in a static scene, offering users real-time feedback and greater control over camera movement. This interactive feature enhances usability, allowing precise trajectory control even for users without specialized knowledge of scene depth. To further improve video quality and control precision, we introduce scene-constrained noise initialization as a mechanism to shape the camera movement in its high-noise stages. By using the preview video of the static 3D scene, RealCam-I2V injects scene-visible regions with controlled noise, guiding the video diffusion model's early generation stages. This high-noise feature constrains the initial layout and camera dynamics, providing a strong foundation for the remaining denoising stages. As denoising progresses, the condition-based approach, trained on metric-scale camera trajectories, preserves global layout and completes the dynamic scene in previously unseen regions. This approach maintains the video diffusion model's capacity for dynamic content generation while ensuring accurate, coherent camera control.

Our experimental results show that RealCam-I2V achieves significant performance gains in video quality and controllability. When relative scales are aligned to metric scales, models such as MotionCtrl, CameraCtrl, and CamI2V see substantial improvements in video quality. Furthermore, with the introduction of scene-constrained noise initialization, RealCam-I2V surpasses state-of-the-art performance benchmarks, particularly on datasets like RealEstate10K [100] and out-of-domain images. These results demonstrate the effectiveness of our approach in both controlled and diverse real-world settings. In summary, our contributions are as follows:

- We identify scale inconsistencies and real-world usability challenges in existing trajectory-based methods and introduce a simple yet effective monocular 3D reconstruction into the preprocessing step of the generation pipeline, serving as a reliable intermediary reference for both training and inference.
- With reconstructed 3D scene, we enable metric-scale training and provide an interactive interface during inference to easily design camera trajectories with preview feedback, along with proposed scene-constrained noise shaping to significantly enhance scene consistency and camera controllability.
- Our method overcomes critical real-world application challenges and achieves substantial improvements on the RealEstate10K dataset, establishing a new sota benchmark both in video quality and control precision.

2. Related Works

Diffusion-based Video Generation. The advancements in diffusion models [52, 53, 94] have led to significant progress in video generation, gradually forming two major categories of foundational tasks: text-to-video and image-to-video. Due to the lack of high-quality video-text datasets [4, 5], previous outstanding works such as AnimateDiff [20], Align Your Latents [5], PYoCo [17], Emu Video [18], LVDM [24], VideoCrafter [9, 10], ModelScope [62],

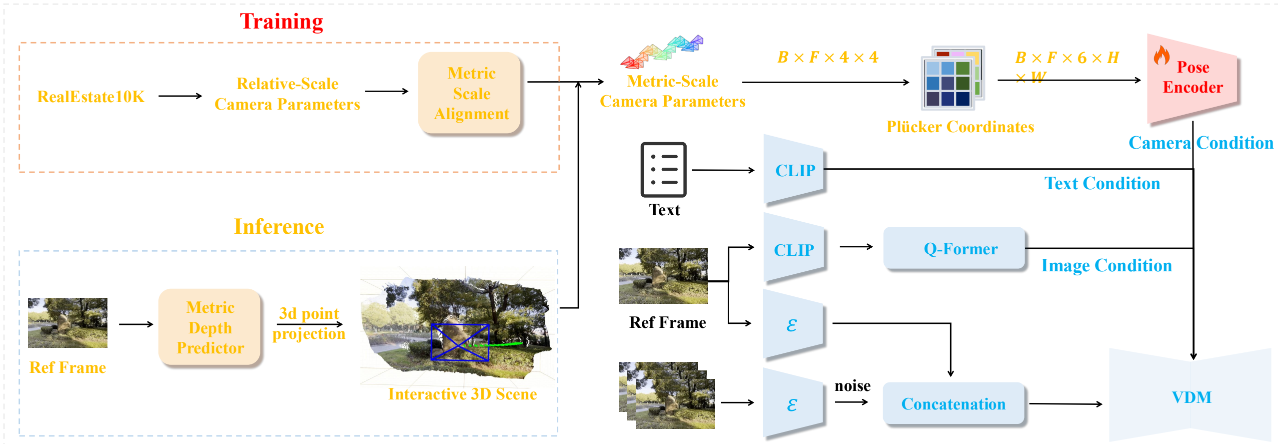

Figure 2. **RealCam-I2V pipeline.** For training, we align camera parameters from relative scale to metric scale. For inference, we use metric depth estimation to construct the point cloud for users to interactively draw the camera trajectory. Due to the metric scale alignment, the user-given camera trajectory in the 3D scene shares the same scene scale as those in real world.

LAVIE [67], and VideoFactory [65] have facilitated text-to-video generation by incorporating Motion Blocks into existing text-to-image [51, 53, 85] models. Building on these T2V efforts, several works introduced images as control signals, leading to a series of I2V models. Representative examples include SVD [4], SEINE [13], PixelDance [87], PIA [92], I2VGen-XL [91], DynamicCrafter [76], and Moonshot [88]. With the improvement in video data [12, 34, 47, 63] and the rapid development of the DiT architecture [42, 49, 86], recent works such as CogVideoX [82], Sora [6], HunyuanVideo [35], GoKu [11], Open-Sora [98], Open-Sora-Plan [40], Vidu [3], Lumina-T2X [16], Vchitect-2.0 [14], RepVideo [55] and Step-Video-T2V[41] have significantly enhanced the quality of video generation. These advancements demonstrate the potential of Video Diffusion Models as realistic world simulators.

Controllable Generation. Controllable generation has become a central focus in both image [32, 45, 50, 57, 72, 75, 83, 90, 95] and video [19, 21, 33, 88] generation, enabling users to direct the output through various types of control. A wide range of controllable inputs has been explored, including text descriptions, pose [27, 43, 64, 79], audio [23, 58, 59], identity representations [7, 37, 68, 73, 74, 93], trajectory [8, 39, 46, 71, 84].

Text-based Camera Control. Text-based camera control methods use natural language descriptions to guide camera motion in video generation. AnimateDiff [20] and SVD [4] fine-tune LoRAs [26] for specific camera movements based on text input. Image conductor[38] proposed to separate different camera and object motions through camera LoRA weight and object LoRA weight to achieve more precise motion control. In contrast, MotionMaster [29] and Peekaboo [31] offer training-free approaches for generating coarse-grained camera motions, though with limited precision. VideoComposer [66] adjusts pixel-level motion vectors to provide finer control, but challenges remain in achieving precise camera control.

Trajectory-based Camera Control. MotionCtrl [69], CameraCtrl [22], and Direct-a-Video [81] use camera pose as input to enhance control, while CVD [36] extends CameraCtrl for multi-view generation, though still limited by motion complexity. To improve geometric consistency, Pose-guided diffusion [60], CamCo [78], and CamI2V [96] apply epipolar constraints for consistent viewpoints. VD3D [2] introduces a ControlNet[90]-like conditioning mechanism with spatiotemporal camera embeddings, enabling more precise control. CamTrol [25] offers a training-free approach that renders static point clouds into multi-view frames for video generation. Cavia [77] introduces view-integrated attention mechanisms to improve viewpoint and temporal consistency, while I2VControl-Camera [15] refines camera movement by employing point trajectories in the camera coordinate system. Recently, 4DiM [70] and AC3D [1] also leverage monocular metric depth estimator to tackling the issues of scale inconsistency. Despite these advancements, challenges in maintaining camera control and scene-scale consistency remain, which our method seeks to address.

3. Method

Our overall pipeline is shown in Fig. 2. For training, we align camera parameters from relative scale to metric scale. For inference, we use metric depth estimation to construct the point cloud for users to interactively draw the camera trajectory. Due to the metric scale alignment, the user-given camera trajectory in the 3D scene shares the same scene scale as those in real world.

3.1. Camera-controlled Image-to-Video Model

Instead of directly modeling the video x, the latent representation $z = \mathcal{E}(x)$ is used for training. The diffusion model ϵ_θ learns to estimate the noise ϵ added at each timestep t, con-

Figure 3. **Scene scale mismatch.** Point clouds reconstructed from metric depth estimation (RGB) are robust and unified, whereas SfM reconstructions (yellow) are relative-scale that may vary across frames. Our alignment enables relative-to-metric conversion of scene scale for real-world applications.

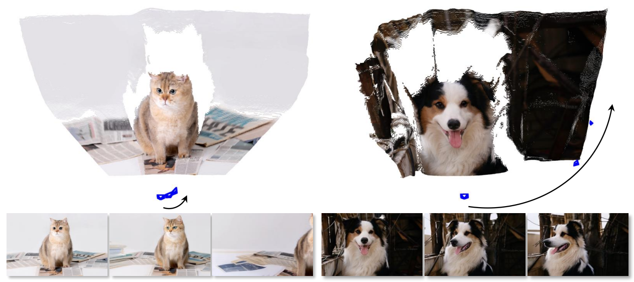

Figure 4. **Camera trajectory ambiguity.** The relative scene-scale measurement fundamentally hinders models from learning physically consistent camera motion.

ditioned on both a text prompt c_{txt}, a reference image c_{img}, and camera condition c_{cam}, with $t \in \mathcal{U}(0,1)$. The training objective simplifies to a reconstruction loss defined as:

$$\mathcal{L} = \mathbb{E}_{z, c_{\text{txt}}, c_{\text{img}}, c_{\text{cam}}, \epsilon, t}\left[\left\|\epsilon - \epsilon_\theta\left(z_t, c_{\text{txt}}, c_{\text{img}}, c_{\text{cam}}, t\right)\right\|_2^2\right], \tag{1}$$

where $z \in \mathbb{R}^{F\times H\times W\times C}$ represents the latent code of a video, with F, H, W, C corresponding to frame count, height, width, and channel dimensions. The noise-corrupted latent code z_t, derived from the ground-truth latent z_0, is expressed as:

$$z_t = \alpha_t z_0 + \sigma_t \epsilon, \tag{2}$$

where $\sigma_t = \sqrt{1-\alpha_t^2}$. Here, α_t and σ_t are hyperparameters governing the diffusion process.

3.2. Metric Scene-scale Alignment

Metric Depth Estimation. To obtain a depth map from a given input image, we use a metric depth predictor f_{depth}, which takes the RGB image I as input and outputs the corresponding depth map $D(u, v)$. The prediction process is formulated as:

$$D(u, v) = f_{\text{depth}}(I),$$

where I is the input RGB image and $D(u, v)$ is the predicted depth value for each pixel at coordinates (u, v). This predicted depth map $D(u, v)$ serves as the foundation for projecting the image into 3D space, allowing us to construct a point cloud in the camera coordinate system. The camera

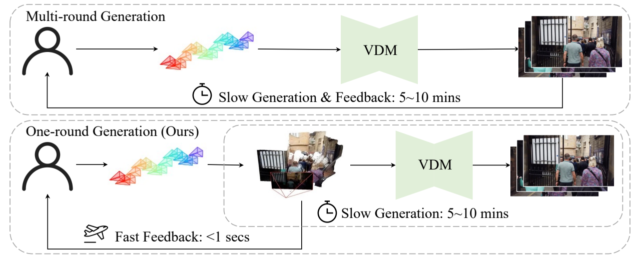

Figure 5. **One-round generation versus multi-round generation.** Our framework decouples camera adjustment via interactive 3D scenes, enabling fast feedback before slow generation.

intrinsics matrix K is defined as:

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix},$$

where f_x and f_y are the focal lengths along the x and y axes, (c_x, c_y) is the principal point of the camera.

Metric Scene-Scale Alignment. To convert camera extrinsics from world-to-camera to an metric-scale camera-to-world representation, we defines that the world-to-camera extrinsics matrix $F_{\text{w2c}} \in \mathbb{R}^{4\times 4}$ is inverted to obtain the corresponding camera-to-world matrix $F_{\text{c2w}} = F_{\text{w2c}}^{-1}$. To express the transformations relative to the first frame, each F_{c2w} is left-multiplied by the camera-to-world matrix of the inverse of first frame $F_{\text{c2w, 1}}$:

$$c_{\text{cam}} = F_{\text{c2w},1}^{-1} \cdot F_{\text{c2w}}.$$

Here, $c_{\text{cam}} \in \mathbb{R}^{F\times 4\times 4}$ represents the camera-to-world transformations aligned relative to the first frame. However, the translation component of c_{cam} remains in a relative scene scale. To convert the relative translation to an metric scale, we align the metric 3D point cloud reconstructed by Depth Anything (metric version) with the 3D point cloud reconstructed by COLMAP (Structure-from-Motion), as shown in Fig. 3 and Fig. 4. The alignment process yields a scale factor α and is applied to the translation component of c_{cam}, resulting in an metric-scale camera-to-world transformation:

$$c_{\text{cam}}^{\text{metric}} = \begin{bmatrix} R & \alpha \cdot T \\ 0 & 1 \end{bmatrix},$$

where R is the rotation matrix, T is the relative translation vector. The resulting $c_{\text{cam}}^{\text{metric}} \in \mathbb{R}^{F\times 4\times 4}$ represents the camera-to-world transformations with metric scene scale, enabling robust and accurate real-world applications.

3.3. One-round Generation in Inference

Users are allowed to draw camera trajectory in the constructed 3D scene and can easily preview the rendering video of expected camera trajectory, free from the costly generation via video diffusion models, as shown in Fig. 5.

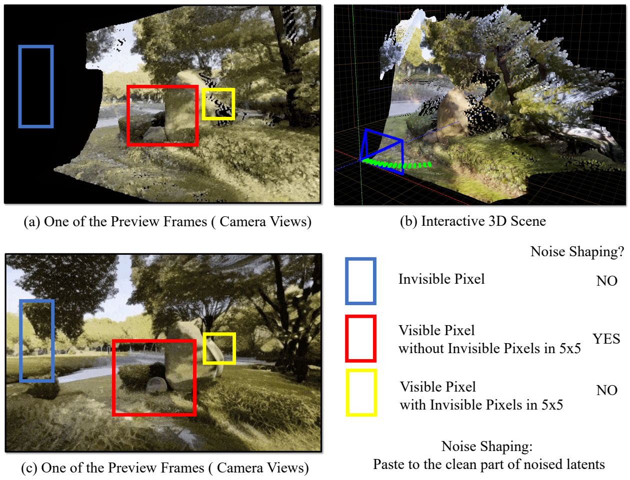

Figure 6. **Pixel selection for noise shaping.** Selected pixels are pasted onto the clean part (predicted z_0) of noised latent z_t. Empirically, applying noise shaping at high noise levels $t > 0.9$ achieves the trade-off of camera control and dynamics.

3D Point Projection For Interaction. Given a depth map $D(u, v)$, the projected 3D coordinates in the camera coordinate system, $\mathbf{p}_c = (x_c, y_c, z_c)^T$, are computed as:

$$\mathbf{p}_c = \mathbf{D}(u, v) \cdot K^{-1} \cdot \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}.$$

Here $(u, v, 1)$ represents the homogeneous coordinates of the pixel, K^{-1} is the inverse of the intrinsic matrix, which maps pixel coordinates to normalized image coordinates. By applying this transformation to all pixels in the depth map, we obtain a set of 3D points $\{\mathbf{p}_c\}$ in the camera coordinate system.

Preview Video Rendering for Camera Motion. With constructed 3D point cloud, we can render a preview video of camera traces. As shown in Fig. 7, the preview video functions as a reference video for validation.

Scene-constrained Noise Shaping. Inspired by previous works SDEdit [44] and DDIM inversion [56], noised features z_t can be used for shaping the layout, camera control of the entire image, especially at timestep with high-level noise. We propose *scene-constrained noise shaping*, which utilizes preview videos generated along user-defined trajectories in the interactive 3D scene. Each frame of the preview video is treated as a reference frame and provided to the generation process during the high-noise stage. The reference frame's pixels are overlaid onto the model-predicted z_0 to achieve the shaping effect.

Next, we detail the process for selecting the pixels to be referenced. As illustrated in Fig. 6, the primary criterion is that a pixel must be visible under the current camera viewpoint in the preview video. To mitigate issues such as holes caused by inaccurate depth predictions, we apply an additional filtering rule: if a visible pixel's $k \times k$ neighborhood contains any invisible pixels, it is considered to lie on an object's edge and potentially affected by depth prediction errors. Such pixels are excluded from selection. Finally, we define the noise shaping process as the formula:

$$z_t = m \cdot (\alpha_t z_{\text{preview}} + \sigma_t \epsilon) + (1 - m) \cdot z_t,$$

where m identifies the selected reference pixels, z_{preview} represents the clean latent features from the preview video. It should be noted that resampling ϵ in each timestep is important because the motivation of noise shaping is to utilize the clean feature of preview feature to shape the low-frequency information of the generation process while the random sampled ϵ may cover some of the valid information in each timestep if it is sampled to be a large value.

Figure 7. **Preview video rendering for user-expected camera trajectory.** It also serves as the reference video in our proposed scene-constrained noise shaping.

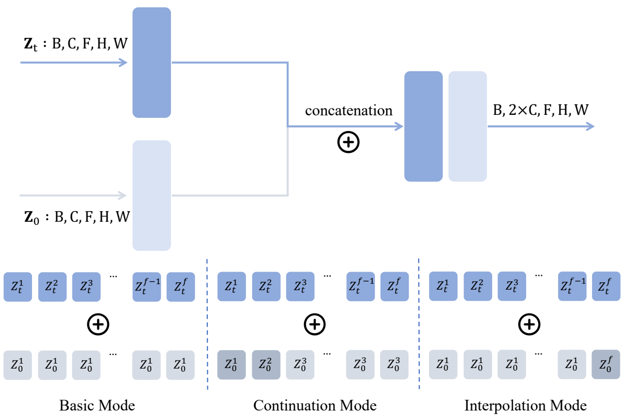

Figure 8. **Concatenation strategies for different tasks.** Basic mode, interpolation mode, and continuation mode can be supported with only minor modification.

3.4. Interpolation, Loop and Continuation

To support different tasks, including interpolation, looping, and continuation for long video generation, we train video diffusion model with different input concatenation mode, as shown in Fig. 8. Given a video latents $z \in \mathbb{R}^{F \times H \times W \times C}$, we define the noised latents of f-th frame at timestep t as z_t^f. We then select i-th clean frame as the condition frame z_0^i. For interpolation mode, we define z_0^{f-1} as the end condition frame. For continuation mode, we define all 1 i-th as condition frame.

Method	RotErr ↓	TransErr ↓		CamMC ↓		FVD ↓	
		Relative Scale	Metric Scale	Relative Scale	Metric Scale	VideoGPT	StyleGAN
DynamiCrafter [76]	3.3415	9.8024	14.135	11.625	15.726	106.02	92.196
MotionCtrl* [69]	1.0527	2.2860	6.8182	2.9312	7.2272	70.292	60.845
CameraCtrl* [22]	0.7373	1.7619	5.5090	2.1644	5.7648	69.202	58.900
CamI2V* [96]	0.4120	1.3409	3.2934	1.5291	3.4230	62.439	53.361
RealCam-I2V (Ours)	**0.3884**	**1.2943**	**2.2317**	**1.4628**	**2.3609**	**53.718**	**45.460**
w.r.t. CamI2V	+5.73%	+3.48%	+32.24%	+4.33%	+31.03%	+13.97%	+14.81%

Table 1. **Quantitative comparison with SOTA methods.** Our approach excels all baselines on both relative and metric results, while coherently improve visual quality of generated videos. We observe over 30% improvement on metric scale results and over 10% improvement on FVD. * denotes our reproduced results on DynamiCrafter [76]. **Best** and second best results are highlighted respectively.

4. Experiments

4.1. Setup

Dataset. We train our model on RealEstate10K [100], which contains 70,000 video clips with well-annotated relative-scale camera poses. For metric scene scale alignment, we run COLMAP [54] point triangulator on each video clip with fixed camera intrinsics and extrinsics directly from RealEstate10K, obtaining the sparse point cloud of the reconstructed scene. For metric depth predictor, we choose Depth Anything V2 [80] Large Indoor, which is fine-tuned on metric depth estimation. We then calculate per-point relative-to-metric scaling factor against metric depth prediction. We term the median value of scaling factors in a frame as the frame-level factor for robustness. To make stable training, we eliminate outliers of video clips whose maximum frame-level scaling factors are among the first 2% with too small values or the last 2% with too large values, assuming sorted in ascending order. The same quantile filtering strategy is applied on the minimum frame-level scaling factors of video clips. It remains 58,000 video clips for training and another 6,000 for testing.

Implementation Details. We choose DynamiCrafter [76] as our image-to-video (I2V) base model and seamlessly integrate proposed RealCam-I2V into it as a plugin. During depth-aligned training, we freeze all parameters of the base model and the depth predictor, while only parameters of proposed method are trainable. We follow DynamiCrafter to sample 16 frames from each single video clip while perform resizing, keeping the aspect ratio, and center cropping to fit in our training scheme. We train the model with a random frame stride ranging from 1 to 10 and take random condition frame as data augmentation. We fix the frame stride to 8 and always use the first frame as the condition frame for inference. We supervise ϵ-prediction on the model of 256×256 resolution and v-prediction on the model of 512×320 resolution respectively, following the pre-training scheme of DynamiCrafter. We apply the Adam optimizer with a constant learning rate of 1×10^{-4} with mixed-precision fp16 and ZeRO-1.

4.2. Metrics

We follow previous works [22, 69, 78, 96] to evaluate camera-controllablity by RotErr, TransErr and CamMC on their estimated camera poses using structure-from-motion (SfM) methods, *e.g.* COLMAP [54] and GLOMAP [48]. We convert the camera pose of each frame in a video clip to be relative to the first frame as canonicalization. We denote the i-th frame relative camera-to-world matrix of ground truth as $\{R_i^{3\times3}, T_i^{3\times1}\}$, and that of generated video as $\{\tilde{R}_i^{3\times3}, \tilde{T}_i^{3\times1}\}$. We randomly select 1,000 samples from test set for evaluation. We sum up per-frame errors as the scene-level result for camera metrics. Inspired by Zheng et al. [96], we repetitively conduct 5 individual trials on each video clips for camera-control metrics to reduce the randomness introduced by SfM tools. Metrics of one video clip are averaged on successful trials at first for later sample-wise average to get final results.

RotError [22, 78, 96]**.** We calculate camera rotation errors by the relative angle between generated videos and ground truths in radians for rotation accuracy.

$$\text{RotErr} = \sum_{i=1}^{n} \arccos \frac{\text{tr}(\tilde{R}_i R_i^{\mathrm{T}}) - 1}{2} \tag{3}$$

TransError [22, 78, 96]**.** We normalize the camera positions by the scene scale of generated video $\tilde{s}_i$ and grouth truth s_i individually for relative TransErr. The scene scale is calculated as the $\mathcal{L}_2$ distance from the first camera to the farthest one for each video clip. For metric TransErr, we normalize both by ground truth video, *i.e.* $\tilde{s}_i = s_i$.

$$\text{TransErr} = \sum_{i=1}^{n} \left\| \frac{\tilde{T}_i}{\tilde{s}_i} - \frac{T_i}{s_i} \right\|_2 \tag{4}$$

CamMC [69]**.** We perform the same normalization strategy as TransError, and evaluate the overall camera pose accuracy by directly calculating the $\mathcal{L}_2$ similarity.

$$\text{CamMC} = \sum_{i=1}^{n} \left\| \left[\tilde{R}_i \,\middle|\, \frac{\tilde{T}_i}{\tilde{s}_i} \right]^{3\times4} - \left[R_i \,\middle|\, \frac{T_i}{s_i} \right]^{3\times4} \right\|_2 \tag{5}$$

Method	MSA	SNS	RotErr ↓	TransErr ↓		CamMC ↓		FVD ↓	
				Relative Scale	Metric Scale	Relative Scale	Metric Scale	VideoGPT	StyleGAN
DynamiCrafter [76]			3.3415	9.8024	14.135	11.625	15.726	106.02	92.196
		✓	1.5163	6.6392	8.4607	7.2108	8.9505	71.942	65.014
MotionCtrl* [69]			1.0527	2.2860	6.8182	2.9312	7.2272	70.292	60.845
	✓		0.8655	2.3342	4.2218	2.8083	4.5984	67.130	58.311
	✓	✓	0.6373	2.0725	3.2308	2.3771	3.4721	58.885	50.111
CameraCtrl* [22]			0.7373	1.7619	5.5090	2.1644	5.7648	69.202	58.900
	✓		0.7042	1.9477	3.8218	2.3007	4.0829	60.314	51.918
	✓	✓	0.5436	1.7954	3.1845	2.0336	3.3620	55.004	46.702
CamI2V* [96]			0.4120	1.3409	3.2934	1.5291	3.4230	62.439	53.361
	✓		0.4243	1.3596	2.6524	1.5539	2.8008	60.516	51.581
RealCam-I2V (Ours)	✓	✓	**0.3884**	**1.2943**	**2.2317**	**1.4628**	**2.3609**	**53.718**	**45.460**

Table 2. **Ablation study of RealCam-I2V plugins.** Metric Scene-scale Alignment (MSA) mitigates scale inconsistency for real-world applications, indicating a more stable and unified camera control. Scene-constrained Noise Shaping (SNS) solely provides substantial improvements on the base model but is less effective than the combined approach (ours). * denotes our reproduced results on DynamiCrafter [76]. **Best** and second best results are highlighted respectively.

Method	Total Score	I2V Subject	I2V Background	Camera Motion	Subject Consistency	Background Consistency	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality
DynamiCrafter [76]	84.22	95.93	**95.43**	22.67	**95.44**	**98.54**	**98.08**	34.15	58.98	62.35
RealCam-I2V (Ours)	**85.71**	**96.14**	95.27	**93.32**	93.96	97.58	97.66	35.77	**59.79**	**63.08**
w.o. SNS	84.37	94.73	93.44	87.42	90.99	96.23	97.36	**46.75**	58.37	62.91
w.o. MSA	83.77	94.71	92.65	91.48	91.24	96.06	97.35	36.99	58.32	63.03

Table 3. **Ablation study on Vbench-I2V [30]**, investigating how camera-conditioned fine-tuning exclusively on static RealEstate10K [100] data affects the generation quality and camera motion of the base model. Notably, even fine-tuned on a static dataset, it preserves dynamics (Dynamic Degree) without compromise. Introducing dynamic datasets will better enhance dynamics, we leave it for future work.

FVD [61]. We also assess the visual quality of generative videos by the distribution distance between generated videos and ground truths.

4.3. Comparison with SOTA Methods

We compare RealCam-I2V with models that either lack camera-condition training (DynamiCrafter [76]) or incorporate camera-condition training, namely MotionCtrl [69] (3×4 camera extrinsics), CameraCtrl [22] (Plücker embedding constructed from camera extrinsics and intrinsics as side input), and CamI2V [96] (current SOTA using Plücker embedding and epipolar attention between all frames), as shown in Tab. 1. We report all results using DynamiCrafter as the base model.

Our method demonstrates significant improvements in visual quality (FVD) and camera control metrics (TransErr, RotErr, CamMC), with particularly over 30% gains on metric scale results. However, these improvements are not fully captured by the RealEstate10K dataset, which contains mostly static scenes.

4.4. Ablation Study

To validate our proposed method, we conducted a comprehensive ablation study on Metric Scene-scale Alignment (MSA) and Scene-constrained Noise Shaping (SNS).

Effect of MSA only. As shown in Tab. 2, compared to relative scale training, metric scale training mitigates scale ambiguity in Fig. 4 and yields notable improvements, especially on metric scale results. It implies that models trained on metric-scale data can more accurately capture true-to-scale translations and better understand camera rotations within a realistic spatial framework. The metric scene scale enhances robustness and compatibility, ensuring that the framework adapts effectively to real-world input and applications. This approach allows for interaction within a unified scale, enabling intuitive user control over camera actions.

Effect of MSA together with SNS. Adding scene-constrained noise shaping to a model with metric scale training yields substantial gains in video quality and camera controllability. This improvement is evident in both camera metrics and FVD. The synergy of metric-scale training and noise shaping ensures robust and precise control in diverse scenarios. This combined approach delivers noticeably better dynamics compared to using noise shaping alone. Large camera movements, rotations, and rapid transitions, which previously struggled to maintain consistency and realism, now work seamlessly. This improvement underscores the strength of integrating metric scale training with noise shaping for complex motion scenarios.

Effect of SNS only. As shown in Tab. 2, scene-constrained noise shaping can be used as the only method for camera

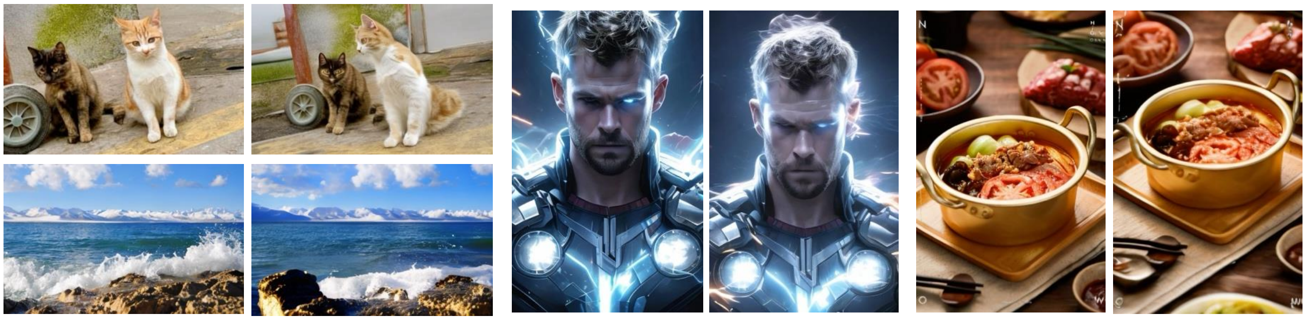

Figure 9. **Visualization on various domains in real life scenarios.** Despite training on RealEstate10K [100], our method can generalize naturally to out-of-domain images, including pets, landscape, anime, food and etc.

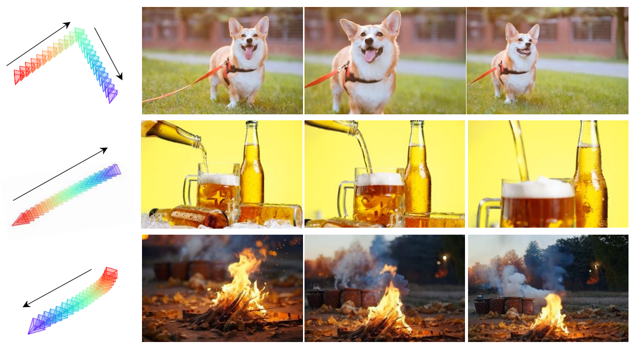

Figure 10. **Visualization on complex trajectory, large movement and video dynamism.** Our framework achieves precise trajectory adherence for complex camera motion paths while preserving high-fidelity video generation with dynamics.

control when applied to a base model not trained with any camera conditions. It provides nearly 50% reduction on DynamiCrafter. However, this method performs poorly compared to the combined method with metric scale training. It also introduces challenges in parameter selection: Applying noise shaping only in the high noise stages limits camera control in lower noise stages, while extending noise shaping to mid-noise phase can suppress video dynamics, resulting in static video output. This limitation affects the fluidity and responsiveness of camera movements, making our combined approach preferable for applications requiring natural dynamics.

4.5. Application

As illustrated in Fig. 9 and Fig. 10, we demonstrate the versatility of our method through visualization results across various applications. Additionally, our results include camera-controlled loop video generation, generative frame interpolation, and smooth scene transitions, highlighting the robustness of our approach. These visualizations showcase two major breakthroughs: first, our method achieves a real-world application breakthrough by addressing challenges like training-inference scale inconsistency and low usability, ensuring improved robustness and compatibility with real-world images. Second, our framework exhibits superior performance in complex camera motions, handling large and rapid movements, rotations, and dynamics more effectively than existing methods.

5. Limitation Analysis and Future Work

Despite training in static RealEstate10K, as discussed in Fig. 9 and Fig. 10, our method demonstrate well generalization to videos with dynamic scenes, various styles and domains due to the designed algorithm to best preserve the knowledge of foundation model, However, the current model still suffers from real-world application mainly due to the limitation of data and there's no suitable dataset designed for camera-controlled video generation, which requires both dynamic scenes and camera movement.

6. Conclusion

In this paper, we address the scale inconsistencies and real-world usability challenges in existing trajectory-based camera-controlled image-to-video generation methods. We introduce a simple yet effective monocular 3D reconstruction into the preprocessing step of the generation pipeline, serving as a reliable intermediary reference for both training and inference. With reconstructed 3D scene, we enable absolute-scale training and provide an interactive interface during inference to easily design camera trajectories with preview feedback, along with proposed scene-constrained noise shaping to significantly enhance scene consistency and camera controllability. Our method overcomes critical real-world application challenges and achieves substantial improvements on the RealEstate10K dataset, establishing a new sota both in video quality and control precision.

7. Acknowledgement

This work is supported in part by National Science Foundation for Distinguished Young Scholars under Grant 62225605, Zhejiang Provincial Natural Science Foundation of China under Grant LD24F020016, Project 12326608 supported by NSFC, Ningbo Science and Technology Special Projects under Grant No. 2025Z028 and the Fundamental Research Funds for the Central Universities.

References

[1] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. *arXiv preprint arXiv:2411.18673*, 2024. 3

[2] Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. *arXiv preprint arXiv:2407.12781*, 2024. 3

[3] Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. *arXiv preprint arXiv:2405.04233*, 2024. 3

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2, 3

[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, pages 22563–22575, 2023. 2

[6] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 3

[7] Hila Chefer, Shiran Zada, Roni Paiss, Ariel Ephrat, Omer Tov, Michael Rubinstein, Lior Wolf, Tali Dekel, Tomer Michaeli, and Inbar Mosseri. Still-moving: Customized video generation without customized video data. *arXiv preprint arXiv:2407.08674*, 2024. 3

[8] Changgu Chen, Junwei Shu, Lianggangxu Chen, Gaoqi He, Changbo Wang, and Yang Li. Motion-zero: Zero-shot moving object control framework for diffusion-based video generation. *arXiv preprint arXiv:2401.10150*, 2024. 3

[9] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 2

[10] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *CVPR*, pages 7310–7320, 2024. 2

[11] Shoufa Chen, Chongjian Ge, Yuqi Zhang, Yida Zhang, Fengda Zhu, Hao Yang, Hongxiang Hao, Hui Wu, Zhichao Lai, Yifei Hu, et al. Goku: Flow based video generative foundation models. *arXiv preprint arXiv:2502.04896*, 2025. 3

[12] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024. 3

[13] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *The Twelfth International Conference on Learning Representations*, 2023. 3

[14] Weichen Fan, Chenyang Si, Junhao Song, Zhenyu Yang, Yinan He, Long Zhuo, Ziqi Huang, Ziyue Dong, Jingwen He, Dongwei Pan, et al. Vchitect-2.0: Parallel transformer for scaling up video diffusion models. *arXiv preprint arXiv:2501.08453*, 2025. 3

[15] Wanquan Feng, Jiawei Liu, Pengqi Tu, Tianhao Qi, Mingzhen Sun, Tianxiang Ma, Songtao Zhao, Siyu Zhou, and Qian He. I2vcontrol-camera: Precise video camera control with adjustable motion strength. 2024. 3

[16] Peng Gao, Le Zhuo, Dongyang Liu, Ruoyi Du, Xu Luo, Longtian Qiu, Yuhang Zhang, Chen Lin, Rongjie Huang, Shijie Geng, et al. Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. *arXiv preprint arXiv:2405.05945*, 2024. 3

[17] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22930–22941, 2023. 2

[18] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023. 2

[19] Litong Gong, Yiran Zhu, Weijie Li, Xiaoyang Kang, Biao Wang, Tiezheng Ge, and Bo Zheng. Atomovideo: High fidelity image-to-video generation. *arXiv preprint arXiv:2403.01800*, 2024. 3

[20] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2, 3

[21] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision*, pages 330–348. Springer, 2025. 3

[22] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 2, 3, 6, 7, 1

[23] Xu He, Qiaochu Huang, Zhensong Zhang, Zhiwei Lin, Zhiyong Wu, Sicheng Yang, Minglei Li, Zhiyi Chen, Song-

cen Xu, and Xiaofei Wu. Co-speech gesture video generation via motion-decoupled diffusion model. In *CVPR*, pages 2263–2273, 2024. 3
[24] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022. 2
[25] Chen Hou, Guoqiang Wei, Yan Zeng, and Zhibo Chen. Training-free camera control for video generation. *arXiv preprint arXiv:2406.10126*, 2024. 3
[26] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3
[27] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR*, pages 8153–8163, 2024. 3
[28] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2
[29] Teng Hu, Jiangning Zhang, Ran Yi, Yating Wang, Hongrui Huang, Jieyu Weng, Yabiao Wang, and Lizhuang Ma. Motionmaster: Training-free camera motion transfer for video generation. *arXiv preprint arXiv:2404.15789*, 2024. 2, 3
[30] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024. 7
[31] Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Behl. Peekaboo: Interactive video generation via masked-diffusion. In *CVPR*, pages 8079–8088, 2024. 2, 3
[32] Rui Jiang, Guang-Cong Zheng, Teng Li, Tian-Rui Yang, Jing-Dong Wang, and Xi Li. A survey of multimodal controllable diffusion models. *Journal of Computer Science and Technology*, 39(3):509–541, 2024. 3
[33] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *CVPR*, pages 6689–6700, 2024. 3
[34] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *Advances in Neural Information Processing Systems*, 37:48955–48970, 2024. 3
[35] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 3
[36] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas Guibas, and Gordon Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. *arXiv preprint arXiv:2405.17414*, 2024. 3
[37] Hengjia Li, Haonan Qiu, Shiwei Zhang, Xiang Wang, Yujie Wei, Zekun Li, Yingya Zhang, Boxi Wu, and Deng Cai. Personalvideo: High id-fidelity video customization without dynamic and semantic degradation. *arXiv preprint arXiv:2411.17048*, 2024. 3
[38] Yaowei Li, Xintao Wang, Zhaoyang Zhang, Zhouxia Wang, Ziyang Yuan, Liangbin Xie, Yuexian Zou, and Ying Shan. Image conductor: Precision control for interactive video synthesis. *arXiv preprint arXiv:2406.15339*, 2024. 2, 3
[39] Zhengqi Li, Richard Tucker, Noah Snavely, and Aleksander Holynski. Generative image dynamics. In *CVPR*, pages 24142–24153, 2024. 3
[40] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024. 3
[41] Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, et al. Step-video-t2v technical report: The practice, challenges, and future of video foundation model. *arXiv preprint arXiv:2502.10248*, 2025. 3
[42] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024. 3
[43] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4117–4125, 2024. 3
[44] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 5
[45] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 3
[46] Koichi Namekata, Sherwin Bahmani, Ziyi Wu, Yash Kant, Igor Gilitschenski, and David B Lindell. Sg-i2v: Self-guided trajectory control in image-to-video generation. *arXiv preprint arXiv:2411.04989*, 2024. 3
[47] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024. 3
[48] Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes Lutz Schönberger. Global Structure-from-Motion Revisited. In *European Conference on Computer Vision (ECCV)*, 2024. 6, 2

[49] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3
[50] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024. 3
[51] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3
[52] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2
[53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3
[54] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 2, 6
[55] Chenyang Si, Weichen Fan, Zhengyao Lv, Ziqi Huang, Yu Qiao, and Ziwei Liu. Repvideo: Rethinking cross-layer representation for video generation. *arXiv preprint arXiv:2501.08994*, 2025. 3
[56] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5
[57] Kunpeng Song, Yizhe Zhu, Bingchen Liu, Qing Yan, Ahmed Elgammal, and Xiao Yang. Moma: Multimodal llm adapter for fast personalized image generation. *arXiv preprint arXiv:2404.05674*, 2024. 3
[58] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3
[59] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. *arXiv preprint arXiv:2402.17485*, 2024. 3
[60] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *CVPR*, pages 16773–16783, 2023. 3
[61] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 7
[62] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 2
[63] Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. *arXiv preprint arXiv:2410.08260*, 2024. 3
[64] Tan Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for referring human dance generation in real world. *arXiv e-prints*, pages arXiv–2307, 2023. 3
[65] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation, 2024. 3
[66] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3
[67] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 3
[68] Zhao Wang, Aoxue Li, Enze Xie, Lingting Zhu, Yong Guo, Qi Dou, and Zhenguo Li. Customvideo: Customizing text-to-video generation with multiple subjects. *arXiv preprint arXiv:2401.09962*, 2024. 3
[69] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2, 3, 6, 7, 1
[70] Daniel Watson, Saurabh Saxena, Lala Li, Andrea Tagliasacchi, and David J Fleet. Controlling space and time with diffusion models. *arXiv preprint arXiv:2407.07860*, 2024. 3
[71] Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation. *arXiv preprint arXiv:2406.17758*, 2024. 3
[72] Tao Wu, Xuewei Li, Zhongang Qi, Di Hu, Xintao Wang, Ying Shan, and Xi Li. Spherediffusion: Spherical geometry-aware distortion resilient diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6126–6134, 2024. 3
[73] Tao Wu, Yong Zhang, Xiaodong Cun, Zhongang Qi, Junfu Pu, Huanzhang Dou, Guangcong Zheng, Ying Shan, and Xi Li. Videomaker: Zero-shot customized video generation with the inherent force of video diffusion models. *arXiv preprint arXiv:2412.19645*, 2024. 3
[74] Tao Wu, Yong Zhang, Xintao Wang, Xianpan Zhou, Guangcong Zheng, Zhongang Qi, Ying Shan, and Xi Li. Customcrafter: Customized video generation with preserving motion and concept composition abilities. *arXiv preprint arXiv:2408.13239*, 2024. 3
[75] Yinwei Wu, Xianpan Zhou, Bing Ma, Xuefeng Su, Kai Ma, and Xinchao Wang. Ifadapter: Instance feature con-

trol for grounded text-to-image generation. *arXiv preprint arXiv:2409.08240*, 2024. 3

[76] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023. 3, 6, 7, 1

[77] Dejia Xu, Yifan Jiang, Chen Huang, Liangchen Song, Thorsten Gernoth, Liangliang Cao, Zhangyang Wang, and Hao Tang. Cavia: Camera-controllable multi-view video diffusion with view-integrated attention. *arXiv preprint arXiv:2410.10774*, 2024. 3

[78] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024. 2, 3, 6

[79] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *CVPR*, pages 1481–1490, 2024. 3

[80] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 2, 6, 1

[81] Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 3

[82] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2, 3

[83] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3

[84] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 3

[85] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 3

[86] Sihyun Yu, Weili Nie, De-An Huang, Boyi Li, Jinwoo Shin, and Anima Anandkumar. Efficient video diffusion models via content-frame motion-latent decomposition. *arXiv preprint arXiv:2403.14148*, 2024. 3

[87] Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. In *CVPR*, pages 8850–8860, 2024. 3

[88] David Junhao Zhang, Dongxu Li, Hung Le, Mike Zheng Shou, Caiming Xiong, and Doyen Sahoo. Moonshot: Towards controllable video generation and editing with multimodal conditions. *arXiv preprint arXiv:2401.01827*, 2024. 3

[89] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arxiv:2410.03825*, 2024. 2

[90] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3

[91] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 3

[92] Yiming Zhang, Zhening Xing, Yanhong Zeng, Youqing Fang, and Kai Chen. Pia: Your personalized image animator via plug-and-play modules in text-to-image models. In *CVPR*, pages 7747–7756, 2024. 3

[93] Ruisi Zhao, Zechuan Zhang, Zongxin Yang, and Yi Yang. 3d object manipulation in a single image using generative models. *arXiv preprint arXiv:2501.12935*, 2025. 3

[94] Guangcong Zheng, Shengming Li, Hui Wang, Taiping Yao, Yang Chen, Shouhong Ding, and Xi Li. Entropy-driven sampling and training scheme for conditional diffusion generation. In *European Conference on Computer Vision*, pages 754–769. Springer, 2022. 2

[95] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *CVPR*, pages 22490–22499, 2023. 3

[96] Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024. 2, 3, 6, 7, 1

[97] Guangcong Zheng, Teng Li, Xianpan Zhou, and Xi Li. Realcam-vid: High-resolution video dataset with dynamic scenes and metric-scale camera movements. *arXiv preprint arXiv:2504.08212*, 2025. 2

[98] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024. 3

[99] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018. 2

[100] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *SIGGRAPH*, 2018. 2, 6, 7, 8, 1