

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

Teng Li, Guangcong Zheng, Rui Jiang, Shuigenzhan, Tao Wu, Yehao Lu, Yining Lin, Xi Li

College of Computer Science & Technology, Zhejiang University

{guangcongzheng, xilizju}@zju.edu.cn

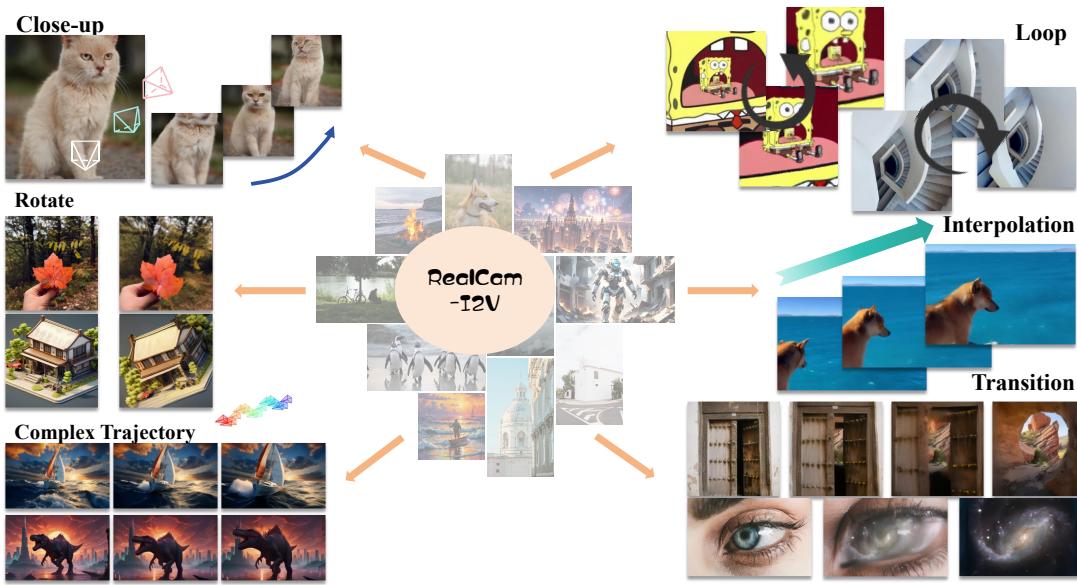


Figure 1. We propose RealCam-I2V, a camera controllable image-to-video generation framework for complex real-world camera control and extra applications including camera-controlled loop video generation, generative frame interpolation, and smooth scene transitions.

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 absolute values, 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, 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. We will release our absolute-scale annotation, codes, and all checkpoints. Please see dynamic results in <https://zgctroy.github.io/RealCam-I2V>.

1. Introduction

Recent advancements in image-to-video generation [4, 7, 8, 14, 18, 66] have significantly improved controllability over synthesized videos. However, challenges remain in achieving realistic, controllable camera movement within com-

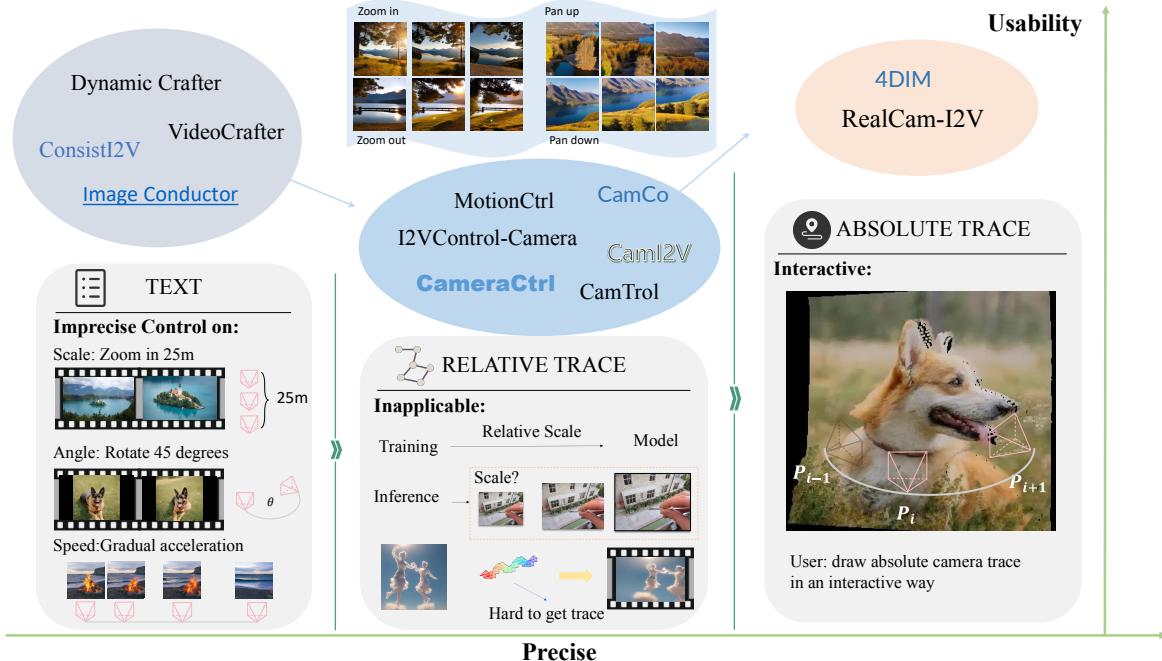


Figure 2. Comparision between text, relative trajectory, and absolute trajectory based camera-controlled image-to-video generation methods on aspects of camera control precision and usability.

plex real-world scenes. Text-based camera-control methods [2, 14, 22, 24, 28, 50], 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 [53], CameraCtrl [16], CamCo [62], and CamI2V [77]. 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 absolute 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, absolute-scale 3D scene. Our approach leverages the Depth Anything v2 [64] model 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 absolute 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 [40], a structure-from-motion (SfM) method. This alignment allows us to rescale COLMAP-annotated camera parameters to the Depth Anything metric, providing an absolute, stable, and robust scale across training data. By aligning relative-scale camera parameters to absolute 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 generation process 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

absolute-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 absolute 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 [78] 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 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 benchmark both in video quality and control precision.

2. Related Works

Diffusion-based Video Generation. The advancement of diffusion models [38, 39, 75] has led to significant progress in video generation. Due to the scarcity of high-quality video-text datasets [2, 3], researchers have adapted existing text-to-image (T2I) models to facilitate text-to-video (T2V) generation. Notable examples include AnimateDiff [14], Align your Latents [3], PYoCo [11], and Emu Video [12]. Further advancements, such as LVDM [18], VideoCrafter [7, 8], ModelScope [47], LAVIE [51], and VideoFactory [49], have refined these approaches by fine-tuning both spatial and temporal blocks, leveraging T2I models for initialization to improve video quality. Recently, Sora [4] and CogVideoX [66] enhance video generation by introducing Transformer-based diffusion backbones [30, 36, 69] and utilizing 3D-VAE, unlocking the potential for realistic world simulators. Additionally, SVD [2], SEINE [9], PixelDance [70] and PIA [74] have made significant strides in image-to-video generation, achieving notable improve-

ments in quality and flexibility. Further, I2VGen-XL [73], DynamicCrafter [59], and Moonshot [71] incorporate additional cross-attention layers to strengthen conditional signals during generation.

Controllable Generation. Controllable generation has become a central focus in both image [25, 33, 37, 42, 56, 58, 67, 72, 76] and video [13, 15, 26, 71] 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 [21, 31, 48, 63], audio [17, 43, 44], identity representations [5, 52, 57], trajectory [6, 29, 34, 55, 68].

Text-based Camera Control. Text-based camera control methods use natural language descriptions to guide camera motion in video generation. AnimateDiff [14] and SVD [2] fine-tune LoRAs [20] for specific camera movements based on text input. Image conductor[28] 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 [22] and Peekaboo [24] offer training-free approaches for generating coarse-grained camera motions, though with limited precision. VideoComposer [50] adjusts pixel-level motion vectors to provide finer control, but challenges remain in achieving precise camera control.

Trajectory-based Camera Control. MotionCtrl [53], CameraCtrl [16], and Direct-a-Video [65] use camera pose as input to enhance control, while CVD [27] extends CameraCtrl for multi-view generation, though still limited by motion complexity. To improve geometric consistency, Pose-guided diffusion [45], CamCo [61], and CamI2V [77] apply epipolar constraints for consistent viewpoints. VD3D [1] introduces a ControlNet[72]-like conditioning mechanism with spatiotemporal camera embeddings, enabling more precise control. CamTrol [19] offers a training-free approach that renders static point clouds into multi-view frames for video generation. Cavia [60] introduces view-integrated attention mechanisms to improve viewpoint and temporal consistency, while I2VControl-Camera [10] refines camera movement by employing point trajectories in the camera coordinate system. Despite these advancements, challenges in maintaining camera control and scene-scale consistency remain, which our method seeks to address. It is noted that 4Dim [54] introduces absolute scale but in 4D novel view synthesis (NVS) of scenes.

3. Method

3.1. Metric Depth Estimation for 3D Reconstruction

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 $D(u, v)$.

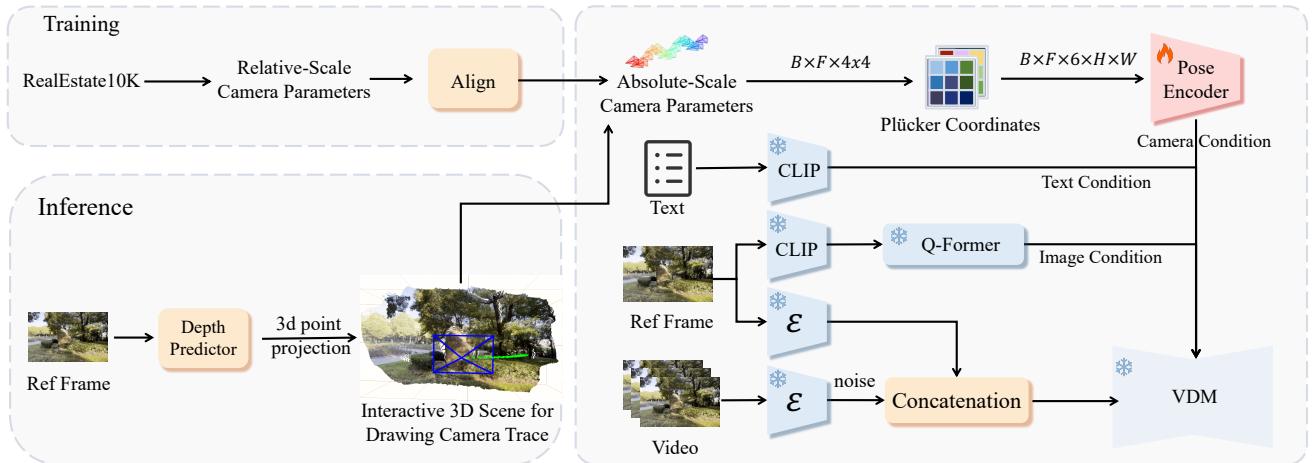


Figure 3. **Pipeline.** In training, we align camera parameters in RealEstate10K from relative scale to absolute scale. In inference, we use metric depth estimation method to construct a 3d point cloud for users to interactively drawing camera traces.



Figure 4. 3D point cloud reconstructed from metric depth estimation (RGB) is robust and unified, whereas the SFM-based reconstruction by methods like COLMAP (Yellow) used in RealEstate10K annotations is in relative scale and may vary across images. Aligning these two 3D scenes enables the transformation from relative to absolute scale (real-world scale).

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 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. 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:

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

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.

3.2. Absolute-Scale Training

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 , conditioned 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:

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

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:

$$z_t = \alpha_t z_0 + \sigma_t \epsilon, \quad (2)$$

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

Aligning from Relative Scale to Absolute Scale. To convert camera extrinsics from world-to-camera to an absolute-scale camera-to-world representation, we define that the world-to-camera extrinsics matrix $F_{w2c} \in \mathbb{R}^{4 \times 4}$ is inverted to obtain the corresponding camera-to-world matrix:

$$F_{c2w} = F_{w2c}^{-1}.$$

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 $F_{c2w,1}$:

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

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 absolute scale, we align the metric 3D point cloud reconstructed by Depth Anything with the 3D point cloud reconstructed by COLMAP (Structure-from-Motion), as shown in Fig. 4. The alignment process yields a scale factor a and is applied to the translation component of c_{cam} , resulting in an absolute-scale camera-to-world transformation:

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

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

3.3. Scene-Constrained Noise Shaping

Inspired by SDEdit [32] and DDIM [41] inversion, 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. 5, 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

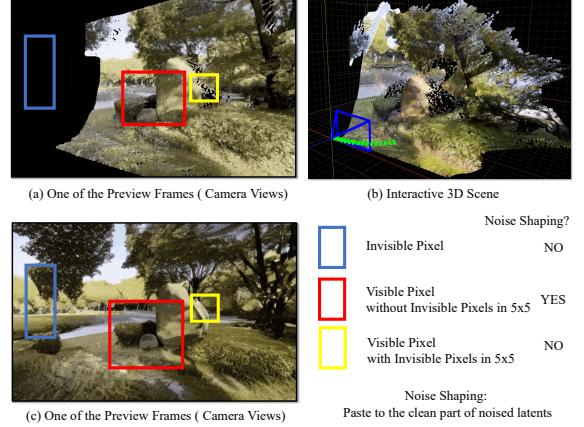


Figure 5. Pixels selected for scene-constrained noise shaping, where they are pasted onto the clean part (predicted z_0) of a noised latent z_t , typically at high noise levels $0.9 < t < 1.0$ is enough for camera control and maintain dynamics.

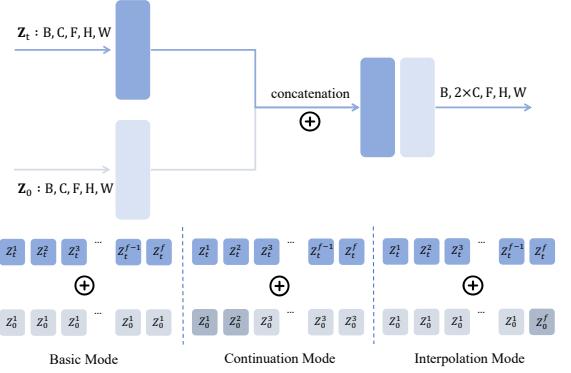


Figure 6. Concatenation for different tasks, including basic mode, interpolation mode, and continuation mode.

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 with the following formula:

$$z_{\text{predict}} = \text{mask} \cdot z_{\text{preview}} + (1 - \text{mask}) \cdot z_{\text{predict}},$$

where the mask identifies the selected reference pixels, z_{preview} represents the latent features from the preview video, and z_{predict} is the model-predicted latent representation.

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. 6. 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.

4. Experiments

4.1. Setup

Dataset. We train our model on RealEstate10K [78], which contains $\sim 70,000$ video clips with well-annotated camera poses. For metric depth alignment of absolute scene scale, we run COLMAP [40] point triangulator on each video clip with fixed camera intrinsics and extrinsics directly from RealEstate10K, obtaining the sparse point cloud of the reconstructed scene. We then calculate per-point depth scale against the metric depth from depth predictor. We term the median value of per-point depth scales in a frame as the frame-level depth scale. To make stable the training, we discard outliers of video clip whose maximum frame-level depth scale of the whole scene is among the top 2% for too small values or the last 2% for too large values, assuming sorted in ascending order. The same quantile filtering strategy is also applied on the minimum frame-level depth scales of video clips. It remains 58,000 video clips for training and another 6,000 for test. During training, 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.

Implementation Details. We choose DynamiCrafter [59] as our image-to-video (I2V) base model and seamlessly integrate proposed RealCam-I2V into it as a plugin. For metric depth predictor, we choose Depth Anything V2 [64] Large Indoor, which is fine-tuned on metric depth estimation. During depth-aligned training, we freeze all parameters of the base model and the depth predictor, while only parameters of proposed method are trainable. More details are listed in dataset section of Appendix. 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 DeepSpeed ZeRO-1. We train proposed method and variants on 8 NVIDIA H100 GPUs with an effective batch size of 64 for 50,000 steps. More details are listed in implementation section of Appendix.

4.2. Metrics

We follow previous works [16, 53, 62, 77] to evaluate camera-controllability by RotErr, TransErr and CamMC on their estimated camera poses using structure-from-motion (SfM) methods, *e.g.* COLMAP [40] and GLOMAP [35]. 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 $\{R_i^{3 \times 3}, T_i^{3 \times 1}\}$, and that of generated video as $\{\tilde{R}_i^{3 \times 3}, \tilde{T}_i^{3 \times 1}\}$. 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. [77], 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. We calculate camera rotation errors by the relative angle between generated videos and ground truths in radians for rotation accuracy.

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

TransError. For relative TransErr, we perform scene scale normalization on the camera positions of each video clip. The scene scale of generated video \tilde{s}_i and ground truth s_i are individually calculated as the \mathcal{L}_2 distance from the first camera to the farthest one for each video clip. For absolute TransErr, we normalize both the video clip to the scene scale of ground truth video, *i.e.* $\tilde{s}_i = s_i$.

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

CamMC. We perform the same scene scale normalization for relative metrics and absolute metrics as TransError, and evaluate the overall camera pose accuracy by directly calculating \mathcal{L}_2 similarity on camera-to-world matrices.

$$\text{CamMC} = \sum_{i=1}^n \left\| \left[\tilde{R}_i \left| \frac{\tilde{T}_i}{\tilde{s}_i} \right. \right]^{3 \times 4} - \left[R_i \left| \frac{T_i}{s_i} \right. \right]^{3 \times 4} \right\|_2 \quad (5)$$

FVD. We also assess the visual quality of generative videos by the distribution distance FVD [46] between generated videos and ground-truths.

4.3. Comparison with SOTA Methods

We compare our proposed method against models that either lack camera-condition training (DynamiCrafter [59]) or incorporate camera-condition training, namely DynamiCrafter+MotionCtrl [53] (3×4 camera extrinsics), DynamiCrafter+CameraCtrl [16] (Plücker embedding constructed

Method	RotErr ↓	TransErr ↓		CamMC ↓		FVD ↓	
		Rel.	Abs.	Rel.	Abs.	VideoGPT	StyleGAN
DynamiCrafter [59]	3.3415	9.8024	14.135	11.625	15.726	106.02	92.196
+ MotionCtrl [53]	1.0366	2.4932	5.8483	3.1244	6.2625	67.253	57.205
+ CameraCtrl [16]*	0.7373	1.7619	5.5090	2.1644	5.7648	69.202	58.900
+ CamI2V [77]*	0.4968	1.4853	3.4069	1.7253	3.5786	63.869	55.276
+ RealCam-I2V (Ours)	0.4052	1.3087	2.4709	1.4869	2.6095	55.229	48.080
w.r.t. CamI2V	+18.44%	+11.89%	+27.47%	+13.82%	+27.08%	+13.53%	+13.02%

Table 1. **Quantitative comparison with SOTA methods.** * denotes the results we reproduced using DynamiCrafter as base I2V model. Our method achieves the state-of-the-art performance on both relative and absolute camera-controllable metrics, while coherently improve visual quality of generated videos, witnessed by a further drop of FVD. **Best** and second best results are highlighted respectively. We observe nearly +30% improvement on absolute metrics while over +10% improvement on relative metrics and FVD.

Method	Total Score	Quality Score						I2V Score		
		subject consistency	background consistency	motion smoothness	dynamic degree	aesthetic quality	imaging quality	i2v subject	i2v background	camera motion
Dynamicrofter (official)	85.25	94.74	98.29	97.83	40.57	58.71	62.28	97.05	97.56	20.92
Dynamicrofter (reproduced)	84.22	95.44	98.54	98.08	34.15	58.98	62.35	95.93	95.43	22.67
+ CamI2V	83.77	91.24	96.06	97.35	36.99	58.32	63.03	94.71	92.65	91.48
++ Absolute Scale	84.37	90.99	96.23	97.36	46.75	58.37	62.91	94.73	93.44	87.42
+++ Noise Shaping (ours)	85.71	93.96	97.58	97.66	35.77	59.79	63.08	96.14	95.27	93.32

Table 2. Evaluation results on Vbench-I2V [23], a widely used benchmark suite with dynamic scenes and various types. Due to the efficient frozen parameters finetuning on Dynamicrofter, our method obtains the ability of camera control but decreases little in other metrics despite only training on static RealEstate10K.

from camera extrinsics and intrinsics as side input), and DynamiCrafter+CamI2V [77] (current SOTA using Plücker embedding and epipolar attention between all frames), as shown in Tab. 1.

Our method demonstrates significant improvements in visual quality (FVD) and camera control metrics (TransErr, RotErr, CamMC), particularly on absolute metrics. Specifically, absolute camera metrics improve by +27%, relative camera metrics by +14%, and FVD by +13%. However, these improvements are not fully captured by the RealEstate10K dataset, which has limitations on movement speed and contains mostly static scenes. It's strongly recommended to view the dynamic visualizations in the supplementary materials for a more comprehensive evaluation.

4.4. Ablation Study

Effect of absolute-scale training only. As shown in Tab. 3, compared to relative-scale training, absolute-scale training yields notable improvements, especially on absolute metrics. It implies that models trained on absolute-scale data can more accurately capture true-to-scale translations and better understand camera rotations within a realistic spatial framework. The absolute scene scale enhances robustness and compatibility, ensuring that the framework adapts effectively to real-world images and applications. This approach allows for interaction within a unified scale, enabling intuitive user control over camera actions.



Figure 7. (a) **Without Scene-Constrained Noise Shaping.** (b) **With Scene-Constrained Noise Shaping.** The left shows failures in large movements without scene-constrained noise shaping, while the right illustrates a loss of dynamics when noise shaping is extended to lower noise stages.

Effect of absolute-scale training + scene-constrained noise shaping. Adding scene-constrained noise shaping to a model trained with absolute-scale yields substantial gains in video quality and camera controllability. This improvement is evident across both camera metrics and FVD. The synergy of absolute-scale training and scene-constrained noise shaping ensures robust and precise control in diverse scenarios. As illustrated in Fig. 7, this combined approach delivers noticeably better dynamics compared to using scene-constrained 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 absolute-scale training with noise

Baseline	RealCam-I2V Plugin			RotErr ↓	TransErr ↓		CamMC ↓		FVD ↓	
	Absolute-Scale	Scene-Constrained	Noise Shaping		Rel.	Abs.	Rel.	Abs.	VideoGPT	StyleGAN
DynamiCrafter* [53]		✓		3.3415 1.5163	9.8024 6.6392	14.135 8.4607	11.625 7.2108	15.726 8.9505	106.02 71.942	92.196 65.014
+ MotionCtrl* [53]	✓			1.0527 0.8655 0.6373	2.2860 2.3342 2.0725	6.8182 4.2218 3.2308	2.9312 2.8083 2.3771	7.2272 4.5984 3.4721	70.292 67.130 58.885	60.845 58.311 50.111
+ CameraCtrl* [16]	✓		✓	0.7373 0.7042 0.5436	1.7619 1.9477 1.7954	5.5090 3.8218 3.1845	2.1644 2.3007 2.0336	5.7648 4.0829 3.3620	69.202 60.314 55.004	58.900 51.918 46.702
+ CamI2V* [77]	✓		✓	0.4968 0.4596 0.4052	1.4853 1.4109 1.3087	3.4069 3.0925 2.4709	1.7253 1.6282 1.4869	3.5786 3.2411 2.6095	63.869 64.451 55.229	55.276 55.313 48.080

Table 3. **Ablation study.** * denotes the results we reproduced using DynamiCrafter as base I2V model. Absolute scene-scale training resolves scale inconsistencies for real-world applications and its improvement on relative metrics indicates a more stable and unified camera control for video generation. Scene-constrained noise shaping can provides substantial improvements in dynamics and large camearabut is less effective than the combined approach, struggling with parameter tuning and dynamic consistency in lower noise stages. **Best** and **second best** results are highlighted respectively.



Figure 8. Without kernel size ≥ 3 in noise shaping, unvisible regions will be wrongly pasted to the generated video.



Figure 9. **Visualization of various applications.** Best viewed as dynamic videos in the supplementary materials.

shaping for complex motion scenarios.

Effect of scene-constrained noise shaping only. As shown in Tab. 3, scene-constrained noise shaping can be used as the sole method for camera control when applied to a base model not trained with any camera conditions. It provides notable improvements in metrics, exemplified by nearly 50% reduction on DynamiCrafter. However, this method underperforms compared to the combined method with absolute-scale training. It also introduces challenges in parameter selection. Applying shaping only in the high-noise phase limits camera contrl in lower noise stages, while extending shaping to mid-noise phase can suppress dynamic elements, resulting in static video output. This limitation affects the fluidity and responsiveness of generated camera movements, making the combination approach preferable for applications requiring natural dynamics.

4.5. Application

As illustrated in Fig. 9, we demonstrate the versatility of our method through visualization results across various applications, including videos generated at resolutions of 512×320 and 1024×576 with camera control under complex scenarios, such as large movements or rotations. 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. More ex-

tensive results are provided in the supplementary materials.

5. Limitation Analysis and Future Work

The model was trained on datasets such as RealEstate10K, which consists primarily of real-world, indoor and outdoor videos collected from YouTube. This dataset's content focuses heavily on realistic, static scenes, resulting in a model that excels in these contexts but performs less effectively when applied to scenes with significantly different visual styles, such as anime, oil paintings, or cartoon-like aesthetics. Better data quality, designing algorithm or improving the ability of fundamental model especially in long video generation will be considered into the future research.

6. Potential Negative Societal Impacts

The image-to-video generation technology developed in this work, with its enhanced camera controllability and breakthrough in real-world applications holds the potential for misuse, particularly in the creation of falsified or deceptive video content. The ability to precisely control camera movements and generate realistic sequences from single images could be exploited to produce convincing yet fabricated videos, leading to ethical concerns around misinformation and privacy violations. To mitigate these risks, we advocate for responsible usage and adherence to ethical guidelines when deploying the RealCam-I2V model.

7. 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.

References

- [1] 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
- [2] 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
- [3] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 3
- [4] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators, 2024. 1, 3
- [5] 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
- [6] 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
- [7] 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. 1, 3
- [8] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024. 1, 3
- [9] 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
- [10] 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. *arXiv preprint arXiv:2411.06525*, 2024. 3
- [11] 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. 3
- [12] 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. 3
- [13] 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
- [14] 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. 1, 2, 3
- [15] 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
- [16] 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, 8, 1
- [17] Xu He, Qiaochu Huang, Zhensong Zhang, Zhiwei Lin, Zhiyong Wu, Sicheng Yang, Minglei Li, Zhiyi Chen, Songcen Xu, and Xiaofei Wu. Co-speech gesture video generation via motion-decoupled diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2263–2273, 2024. 3
- [18] 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. 1, 3
- [19] Chen Hou, Guoqiang Wei, Yan Zeng, and Zhibo Chen. Training-free camera control for video generation. *arXiv preprint arXiv:2406.10126*, 2024. 3

- [20] 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
- [21] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 3
- [22] 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
- [23] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024. 7
- [24] Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Behl. Peekaboo: Interactive video generation via masked-diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8079–8088, 2024. 2, 3
- [25] Rui Jiang, Guang-Cong Zheng, Teng Li, Tian-Rui Yang, Jing-Dong Wang, and Xi Li. A survey of multi-modal controllable diffusion models. *Journal of Computer Science and Technology*, 39(3):509–541, 2024. 3
- [26] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6689–6700, 2024. 3
- [27] 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
- [28] 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
- [29] Zhengqi Li, Richard Tucker, Noah Snavely, and Aleksander Holynski. Generative image dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24142–24153, 2024. 3
- [30] 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
- [31] 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
- [32] 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
- [33] 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
- [34] 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
- [35] Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes Lutz Schönenberger. Global Structure-from-Motion Revisited. In *European Conference on Computer Vision (ECCV)*, 2024. 6
- [36] 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
- [37] 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
- [38] 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. 3
- [39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3
- [40] 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
- [41] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5

- [42] Kunpeng Song, Yizhe Zhu, Bingchen Liu, Qing Yan, Ahmed Elgammal, and Xiao Yang. Moma: Multi-modal llm adapter for fast personalized image generation. *arXiv preprint arXiv:2404.05674*, 2024. 3
- [43] 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
- [44] 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
- [45] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Al-sisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16773–16783, 2023. 3
- [46] 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. 6
- [47] 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. 3
- [48] 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
- [49] Wenjing Wang, Huan Yang, Zixi Tuo, Huigu He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Video-factory: Swap attention in spatiotemporal diffusions for text-to-video generation, 2024. 3
- [50] 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
- [51] 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
- [52] 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
- [53] 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, 8, 1
- [54] 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
- [55] 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
- [56] 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
- [57] 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
- [58] Yinwei Wu, Xianpan Zhou, Bing Ma, Xuefeng Su, Kai Ma, and Xinchao Wang. Ifadapter: Instance feature control for grounded text-to-image generation. *arXiv preprint arXiv:2409.08240*, 2024. 3
- [59] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamiccrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023. 3, 6, 7, 1
- [60] 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
- [61] 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. 3
- [62] 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, 6
- [63] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024. 3
- [64] 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
- [65] 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
- [66] 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. 1, 3
- [67] 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
- [68] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Drag-nuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 3
- [69] 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
- [70] Yan Zeng, Guoqiang Wei, Jian Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8850–8860, 2024. 3
- [71] 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
- [72] 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
- [73] 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
- [74] Yiming Zhang, Zhenning Xing, Yanhong Zeng, Youqing Fang, and Kai Chen. Pia: Your personalized image animator via plug-and-play modules in text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7747–7756, 2024. 3
- [75] 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. 3
- [76] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhonggang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22490–22499, 2023. 3
- [77] 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, 8, 1
- [78] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *SIGGRAPH*, 2018. 3, 6, 1

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

Supplementary Material

We highly recommend that reviewers refer to [index.html](#) provided in our supplementary files.

A. Dataset

The camera trajectory of each video clip from RealEstate10K¹ [78] is first derived by SLAM methods at lower resolution with the field of view fixed at 90°. The authors then refine each of camera sequence using a structure-from-motion (SfM) pipeline, performing feature extraction, feature matching and global bundle adjustment successively. Given the unawareness of global scene scale, the resulted camera poses of RealEstate10K are up to an arbitrary scale per clip. For each frame the authors compute the 5-th percentile depth among all point depths from that frame’s camera. Computing this depth across all cameras in a video clip gives a set of near plane depths and the whole scene is scaled so that the 10-th percentile of this set of depths is 1.25m.

While using RealEstate10K’s scenes and camera trajectories during inference avoids scale issues within the dataset, challenges arise in more general cases. Specifically, when pairing out-of-domain images with either in-domain or out-of-domain trajectories, the inconsistencies between training and inference scales become evident. *These inconsistencies make it impossible to generate realistic and controllable videos.*

The solution lies in reconstructing an absolute-scale scene for any given image. By leveraging metric depth predictor, we can reconstruct the absolute-scale 3D scene for the reference image. This absolute-scale scene bridges the gap between training and inference, enabling robust generalization to real-world applications. With this alignment, the model becomes capable of handling diverse combinations of images and trajectories, ensuring consistent and reliable performance across various scenarios.

B. Depth Predictor

We choose the metric depth version of Depth Anything V2² [64] as the metric depth predictor. Compared to their basic versions, the authors fine-tune the pre-trained encoder on synthetic datasets for indoor and outdoor metric depth estimation. The indoor model is capable of monocular metric depth estimation within a maximum depth of 20m. We



Figure 10. Camera Trajectory Interpolation.

choose Large as the model size, which has 335.3M parameters, and the indoor version. The scene scale of our model is aligned to the metric depth space of Depth Anything V2 Large Indoor, *i.e.* absolute scene scale.

C. Training

We choose DynamiCrafter³ [59] as our image-to-video (I2V) base model. We trained proposed method on 4 publicly accessible variants of DynamiCrafter, namely 256, 512, 512_interp and 1024. We conduct ablation study on resolution 256 × 256, due to the limitation of computing resource. For resolution 256 × 256, we train all models on ϵ -prediction with effective batch size 64 on 8 NVIDIA H100 GPUs for 50,000 steps, taking about 25 hours. For resolution 512 × 320 and 1024 × 576, we train RealCam-I2V on v -prediction while enable `perframe_ae` and `gradient_checkpoint` to reduce peak GPU memory consumption. We apply the Adam optimizer with a constant learning rate of 1×10^{-4} with mixed-precision fp16 and DeepSpeed ZeRO-1.

For MotionCtrl [53] and CameraCtrl [16], we reproduce all results on DynamiCrafter for fair comparison. For CamI2V [77], we implement hard mask epipolar attention and set 2 register tokens, aligned with the original paper. In quantitative comparison and ablation study, we set fixed text image CFG to 7.5 and camera CFG to 1.0.

D. Camera Keyframe Interpolation

In real-world applications, user-provided camera trajectories often consist of a limited number of keyframes (*e.g.*, 4 keyframes). To ensure smooth and continuous motion across the trajectory while adhering to the user’s input, we perform linear interpolation in SE(3) space to expand the trajectory to a higher number of frames (*e.g.*, 16 interpolated frames), as shown in Fig. 10. This step ensures that our model generates consistent and visually coherent videos without compromising the accuracy of user-defined camera movements.

¹<https://google.github.io/realestate10k/>

²https://github.com/DepthAnything/Depth-Anything-V2/tree/main/metric_depth

³<https://github.com/Doubiuu/DynamiCrafter>