CAMI2V: CAMERA-CONTROLLED IMAGE-TO-VIDEO DIFFUSION MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advancements have integrated camera pose as a user-friendly and physics-informed condition in video diffusion models, enabling precise camera control. In this paper, we identify one of the key challenges as effectively modeling noisy cross-frame interactions to enhance geometry consistency and camera controllability. We innovatively associate the quality of a condition with its ability to reduce uncertainty and interpret noisy cross-frame features as a form of noisy condition. Recognizing that noisy conditions provide deterministic information while also introducing randomness and potential misguidance due to added noise, we propose applying epipolar attention to only aggregate features along corresponding epipolar lines, thereby accessing an optimal amount of noisy conditions. Additionally, we address scenarios where epipolar lines disappear, commonly caused by rapid camera movements, dynamic objects, or occlusions, ensuring robust performance in diverse environments. Furthermore, we develop a more robust and reproducible evaluation pipeline to address the inaccuracies and instabilities of existing camera control metrics. Our method achieves a 25.64% improvement in camera controllability on the RealEstate10K dataset without compromising dynamics or generation quality and demonstrates strong generalization to out-of-domain images. Training and inference require only 24GB and 12GB of memory, respectively, for 16-frame sequences at 256×256 resolution. We will release all checkpoints, along with training and evaluation code. Dynamic videos are available for viewing on our supplementary anonymous web page.

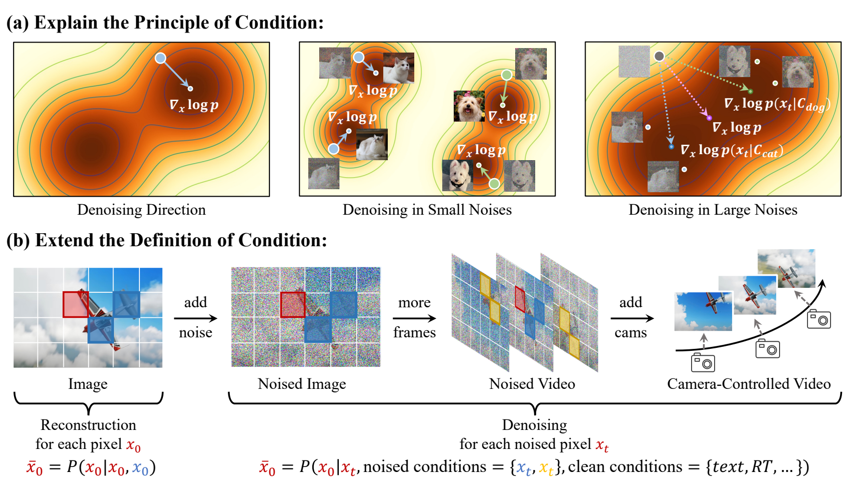

Figure 1: **Rethinking condition in diffusion models.** Diffusion models denoise along the gradient of log probability density function. At large noise levels, the high density region becomes the overlap of numerous noisy samples, resulting in visual blurriness. We point out that *the effectiveness of a condition depends on how much uncertainty it reduces*. From a new perspective, we categorize conditions into *clean conditions* (e.g. texts, camera extrinsics) that remain visible throughout the denoising process, and *noisy conditions* (e.g. noised pixels in the current and other frames) whose deterministic information $\alpha_t x_0$ will be gradually dominated by the randomness of noise $\sigma_t \epsilon$.

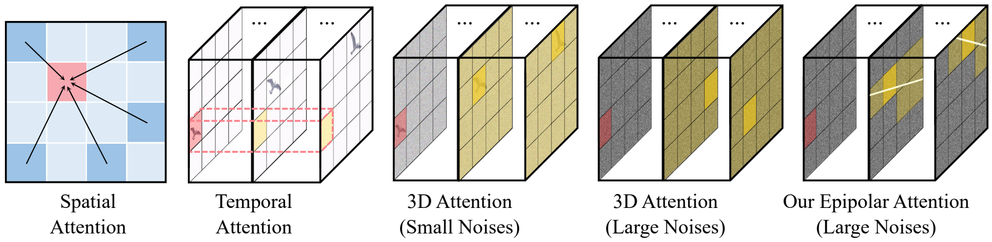

Figure 2: **Comparison of existing attention mechanisms for tracking displaced noised features.** Temporal attention is limited to features at the same location of picture, rendering it ineffective for significant camera movements. In contrast, 3D full attention facilitates cross-frame tracking due to its broad receptive field. However, high noise levels can obscure deterministic information, hindering consistent tracking. Our proposed epipolar attention aggregates features along the epipolar line, effectively modeling cross-frame relationships even under high noise conditions.

1 Introduction

The remarkable 3D consistency demonstrated in videos generated by Sora (Brooks et al., 2024) has highlighted the powerful capabilities of diffusion models (Ho et al., 2020; Rombach et al., 2022), showcasing their potential as a world simulator. Many researchers have attempted to enable the model to understand real-world knowledge (Chen et al., 2023a; Liu et al., 2023).

Condition or guidance (Ho & Salimans, 2022; Dhariwal & Nichol, 2021) is widely recognized as a crucial factor in enhancing generation quality. This is attributed to the fundamental principles that diffusion models denoise along the gradient of the log probability density function (score function) (Song et al., 2020), moving towards a high density region. However, this characteristic has varying effects at different noise levels (Tang et al., 2023a). As shown in Fig. 1(a), the high density region under high noise level becomes the overlap of numerous noisy samples, resulting in visual blurriness. By providing the model with conditions such as c_{dog} and c_{cat}, it can rapidly eliminate incorrect generations. This illustrates that **adding more conditions can guide the model towards desired outcomes while reducing uncertainty.**

Consequently, *incorporating physics-related or more detailed conditions into the diffusion model is an effective way of improving its world understanding*. Considering that video generation requires providing condition for each frame, it is essential to identify a condition that is physics-related but also user-friendly. Recently, some camera-conditioned text-to-video diffusion models such as MotionCtrl (He et al., 2024a) and CameraCtrl (Wang et al., 2024d) have proposed using camera poses of each frame as a new type of condition. However, these methods simply inject camera conditions through a side input (like T2I-Adapter (Mou et al., 2024)) and neglect the inherent physical knowledge of camera pose, resulting in imprecise camera control, inconsistencies, and also poor interpretability.

In this paper, we identify one of the key challenges of camera-controlled image-to-video diffusion models as *how to effectively model noisy cross-frame interactions to enhance geometry consistency and camera controllability.* As illustrated in Fig. 2, separated spatial and temporal attention serves as an indirect form of 3D attention. The cross-frame interaction in temporal attention is confined to features at the same location in the image, rendering it ineffective for tracking significant movements resulting from large camera shifts. 3D full attention is widely applied in advanced video diffusion models such as OpenSora (Zheng et al., 2024) and CogVideoX (Yang et al., 2024b), due to its extensive receptive field. From the novel perspective of the noisy conditions mentioned in Fig. 1, the broad receptive field of 3D full attention allows it to access more noisy conditions. However, we argue that **accessing more noisy conditions does not necessarily reduce uncertainty** and thus not necessarily lead to better performance due to the randomness inherent in the noise. As previously highlighted in Fig. 1, **the quality of a condition is determined by its ability to reduce the model's uncertainty, rather than its quantity**.

To address these issues, we have found that **applying epipolar constraints is one of the most suitable way to prevent the model from being misled by noise**. By restricting attention to features along the epipolar lines, the model can interact with more relevant and less noisy information, improving cross-frame interactions in diffusion models. Specifically, we propose to apply Plücker coordinates (Plücker, 1828) as absolute 3D ray embedding for implicit learning of 3D space and propose a epipolar attention mechanism that introduces an explicit constraint. By doing so, our approach minimizes the search space and reduces potential errors, ultimately enhancing 3D consistency across frames and improving overall controllability. Additionally, inspired by Timothée et al. (2024), we incorporate register tokens into epipolar attention to address scenarios where there are no intersections between frames, often caused by rapid camera movements, dynamic objects, or occlusions.

For inference, we propose a multiple classifier-free guidance scale to control images, text, and camera respectively. If needed, several forward passes can be combined into a single pass by absorbing the scales of image, text, and camera into the model input, similar to timestep conditioning according to (Meng et al., 2023). For evaluation, we identify inaccuracies and instability in the current measurements of camera controllability due to the intrinsic limitations of SfM-based methods such as COLMAP (Schonberger & Frahm, 2016), which rely on identifying keypoint pairs and is quite challenging on generated videos with low resolution, high frame stride, and 3D inconsistencies. Considering the importance of accurate evaluation in this field, we establish a more robust, precise, and reproducible evaluation pipeline by implementing several enhancements. More details are provided in Section 5.

We conduct experiments on the RealEstate10k dataset and evaluate video generation quality using FVD (Unterthiner et al., 2018), as well as camera controllability metrics including RotError, TranError (Wang et al., 2024d), and CamMC (He et al., 2024a). The results demonstrate that the proposed epipolar attention mechanism across all noised frames significantly enhances geometric consistency and improves camera controllability. To facilitate further research, we will release all models trained on open-source frameworks such as DynamiCrafter, along with high-resolution checkpoints and training/evaluation codes, as soon as possible. To summarize, our key contributions are as follows:

- We identify one of the key challenges of camera-controlled image-to-video diffusion models as effectively modeling noisy cross-frame interactions to enhance geometry consistency and camera controllability.
- Well-motivated by the relationship between the quality of a condition and its ability to reduce uncertainty, we innovatively interpret noisy cross-frame features as a form of noisy condition and propose to apply epipolar attention to access an optimal amount of noisy condition. We also address scenarios where epipolar lines disappear by register tokens.
- We point out and analyze the reasons for inaccurate measurement of camera controllability caused by the inherent limitations of SfM evaluator and re-establish a more robust, accurate and reproducible evaluation pipeline. We achieve a 32.96%, 25.64%, 20.77% improvement over CameraCtrl on RotErr, CamMC, TransErr on the RealEstate10K dataset without compromising dynamics, generation quality, or generalization on out-of-domain images.

2 Related Work

Diffusion-based Video Generation. With the advancement of diffusion models (Rombach et al., 2022; Ramesh et al., 2022; Zheng et al., 2022), video generation technology has progressed significantly. Given the scarcity of high-quality video-text datasets (Blattmann et al., 2023a;b), many researchers have sought to adapt existing text-to-image (T2I) models for text-to-video (T2V) generation. Some efforts involve integrating temporal blocks into original T2I models, training these additions to facilitate the conversion to T2V models. Examples include AnimateDiff (Guo et al., 2023), Align your Latents (Blattmann et al., 2023b), PYoCo (Ge et al., 2023), and Emu video (Girdhar et al., 2023). Additionally, methods such as LVDM (He et al., 2022), VideoCrafter (Chen et al., 2023a; 2024b), ModelScope (Wang et al., 2023a), LAVIE (Wang et al., 2023c), and VideoFactory Wang et al. (2024a) have adopted a similar structure, using T2I models as initialization weights and fine-tuning both spatial and temporal blocks to achieve better visual effects. Building on this foundation, Sora (Brooks et al., 2024) and CogVideoX (Yang et al., 2024b) have significantly enhanced video generation capabilities by introducing Transformer-based diffusion backbones (Peebles & Xie, 2023; Ma et al., 2024a; Yu et al., 2024) and leveraging 3D-VAE technology, thereby opening up the possibility of world simulators. Furthermore, works such as Dynamicrafter (Xing et al., 2023), SVD (Blattmann et al., 2023a), Seine (Chen et al., 2023b), I2vgen-XL (Zhang et al., 2023b), and PIA (Zhang et al., 2024) have extensively explored image-to-video generation, achieving substantial progress.

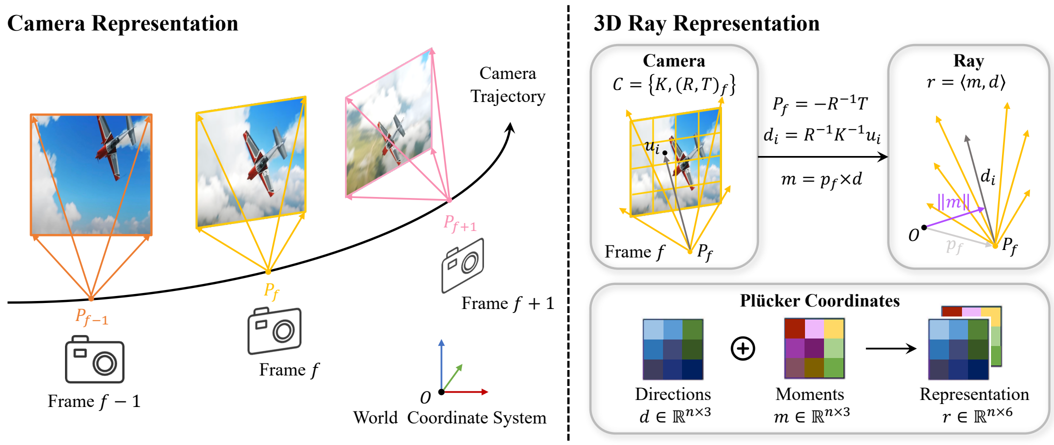

Figure 3: **Parameterizations for cameras.** Left: Camera representation and trajectory visualization in the world coordinate system. Right: The transformation from camera representations to 3D ray representations as Plücker coordinates given pixel coordinates.

Controllable Generation. With the development of image controllable generation technology (Zhang et al., 2023a; Jiang et al., 2024; Mou et al., 2024; Zheng et al., 2023; Peng et al., 2024; Ye et al., 2023; Wu et al., 2024b; Song et al., 2024; Wu et al., 2024d), video controllable generation has gradually become a highly focused direction. Significant progress has been made in areas such as pose (Ma et al., 2024b; Wang et al., 2023b; Hu, 2024; Xu et al., 2024b), trajectory (Yin et al., 2023; Chen et al., 2024a; Li et al., 2024; Wu et al., 2024a), subject (Chefer et al., 2024; Wang et al., 2024c; Wu et al., 2024c), and audio (Tang et al., 2023b; Tian et al., 2024; He et al., 2024b), greatly facilitating users to generate desired videos according to their needs.

Camera-controlled Video Generation. AnimateDiff (Guo et al., 2023) utilizes LoRA (Hu et al., 2021) fine-tuning to achieve specific camera movements. MotionMaster (Hu et al., 2024) and Peekaboo (Jain et al., 2024) explore a training-free method for coarse-grained camera movement generation, but they lack precise control. VideoComposer (Wang et al., 2024b) offers global motion guidance by adjusting pixel-level motion vectors. In contrast, MotionCtrl (Wang et al., 2024d), CameraCtrl (He et al., 2024a), and Direct-a-Video (Yang et al., 2024a) incorporate camera pose information as side input; however, these methods primarily focus on text-to-video generation and do not effectively leverage 3D geometric priors in camera pose. CamCo (Xu et al., 2024a) also facilitates controllable camera generation in the image-to-video task by using epipolar attention (Kant et al., 2024; Tseng et al., 2023) to ensure consistency between generated frames and a single reference frame only. However, it does not account for scenarios where frames lack overlapping regions with the reference frame and can thus be regarded as a degenerate version of our approach.

3 Method

3.1 Preliminaries

3D Ray Embedding. We follow CameraCtrl (He et al., 2024a) to apply plücker embedding as global positional embedding. Considering camera intrinsics $K \in \mathbb{R}^{3\times3}$ and extrinsics (rotation $R \in \mathrm{SO}(3)$, translation $T \in \mathbb{R}^3$), it parameterizes the transform from world coordinates to pixel coordinates by projection $u = K\,[R\,|\,T]\,x$. This low-dimensional representation may hinder neural networks from direct regression. Instead, we follow (Tseng et al., 2023) to represent cameras as ray bundles:

$$\mathcal{R} = \{r_1, \ldots, r_n\}, \tag{1}$$

where each ray $r_i \in \mathbb{R}^6$ is associated with a known pixel coordinate u_i. Each ray $\boldsymbol{r}$ can be parameterized by ray direction $d \in \mathbb{R}^3$ from camera center $P \in \mathbb{R}^3$ as Plücker coordinates:

$$r = \langle m, d \rangle \in \mathbb{R}^6, \tag{2}$$

where $m = p \times d \in \mathbb{R}^3$ is the moment vector. When normalize d to unit length, the norm of the moment m represents the distance from the ray to the world origin. Given a set of 2D pixel

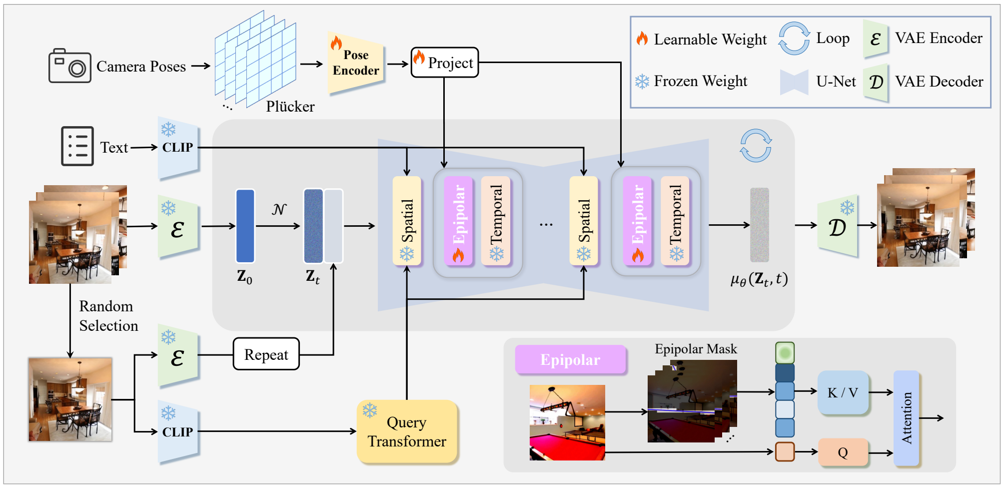

Figure 4: **Pipeline of camera-controlled image-to-video diffusion model.** We follow CameraCtrl to add a learnable pose encoder and a linear projection to process plucker embeddings as a global positional embedding. Epipolar attention is added between spatial and temporal attention.

coordinates $\{(u, v)_i\}^n$, ray directions d can be computed by the unprojection transform:

$$d = R^{-1}K^{-1} \cdot (u, v, 1)^{\mathrm{T}}, \ m = (-R^{-1}T) \times d \tag{3}$$

Text-guided Image to Video Diffusion Model. Text-guided Image to Video Diffusion Model (Zhang et al., 2024; 2023b; Xing et al., 2023) learn a video data distribution by the gradual denoising of a variable sampled from a Gaussian distribution. For image to video generation, first, a learnable auto-encoder (consisting of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$) is trained to compress the video into latent space. Then, a latent representation $z = \mathcal{E}(x)$ is trained instead of a video x. Specifically, the diffusion model ϵ_θ aims to predict the added noise ϵ at each timestep t based on the text condition c_{txt} and the reference image condition c_{img}, where $t \in \mathcal{U}(0, 1)$. The training objective can be simplified as a reconstruction loss:

$$\mathcal{L} = \mathbb{E}_{z, c_{\text{txt}}, c_{\text{img}}, \epsilon \sim \mathcal{N}(0, \mathrm{I}), t} \left[\|\epsilon - \epsilon_\theta \left(\mathbf{z}_t, c_{\text{txt}}, c_{\text{img}}, t \right)\|_2^2 \right], \tag{4}$$

where $\mathbf{z} \in \mathbb{R}^{F \times H \times W \times C}$ is the latent code of video data with F, H, W, C being frame, height, width, and channel. Besides, c_{text} is the text prompt for input video, and c_{img} is the reference frame of video. A noise-corrupted latent code $\mathbf{z}_t$ from the ground-truth z_0 is formulated as $\mathbf{z}_t = \alpha_t z_0 + \sigma_t \epsilon$, where $\sigma_t = \sqrt{1 - \alpha_t^2}$, α_t and σ_t are hyperparameters to control the diffusion process.

3.2 Overall Pipeline

In this section, we present our novel camera-conditioned method for geometry-consistent image-to-video generation, as shown in Fig. 4. We first describe cross-frame epipolar line and discreterized epipolar mask, grounded in the principle of camera projection. Next, we propose epipolar-constrained attention module for the base model in a plug-and-play manner, which effectively make use of feature correlations along epipolar lines. Further, we discuss the situation when epipolar lines of all frames are outside the image plane and introduce register tokens as a simple yet effective fix. Finally, we leverage multiple CFG to balance visual quality and camera pose consistency.

3.3 Epipolar Attention for Noised Features Tracking

Epipolar line and mask. The proposed epipolar attention mechanism seeks to establish a connection between frames, as shown on the left-hand side of Fig. 5. Its primary concept involves utilizing the epipolar line as a constraint, which effectively narrows down the potential matching pixels from one target frame to any other frames. For a single pixel at coordinate (u, v) on the i-th frame, the corresponding epipolar line $l_{ij} \in \mathbb{R}^3$ on the j-th frame can be formulated as:

$$l_{ij}(u, v) = F_{ij} \cdot (u, v, 1)^{\mathrm{T}}, \tag{5}$$

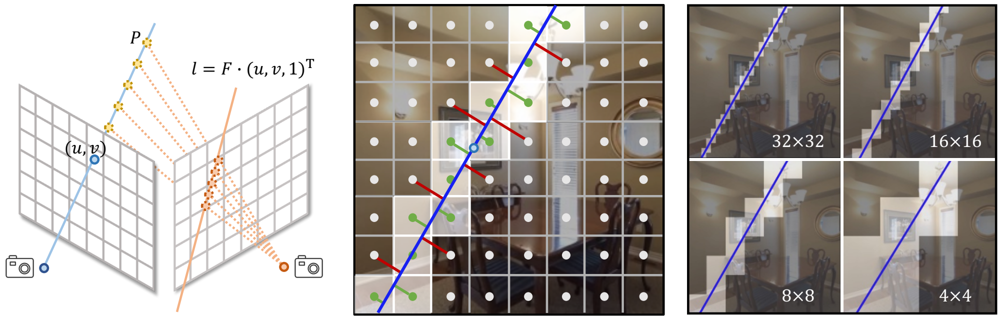

Figure 5: **Epipolar line and mask.** Left: Epipolar constraint of the j-th frame from one pixel at (u, v) on the i-th frame. Middle: Epipolar mask discretized by the distance threshold δ, so that only neighboring pixels in green are allowed to attend while those red lined are not. Right: Multi-resolution epipolar mask adaptive to the feature size in U-Net layers.

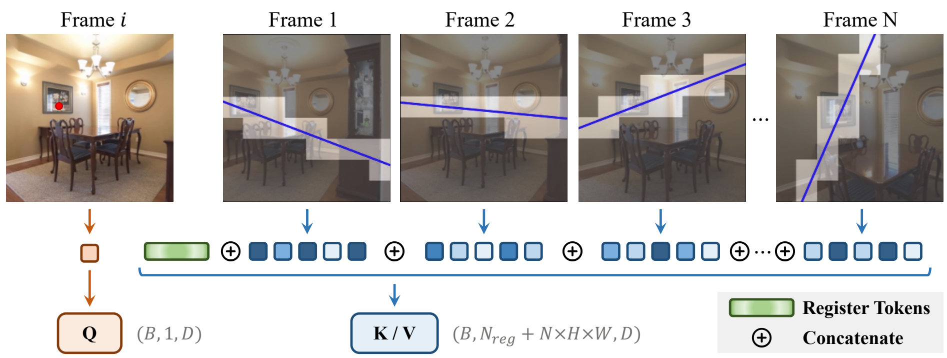

Figure 6: **Epipolar attention mask with register tokens.** We specify query pixel by red point in the i-th frame for clarity. Epipolar attention mask is constructed by concatenating epipolar masks along all frames. We insert register tokens to key/value sequence to deal with zero epipolar scenarios.

where F_{ij} is the camera fundamental matrix of two frames, which can be derived as $F_{ij} = K_j^{-\mathrm{T}} \cdot E_{ij} \cdot K_i^{-1}$ given the camera intrinsics $K_i, K_j \in \mathbb{R}^{3\times3}$ and the camera essential matrix $E_{ij} \in \mathbb{R}^{3\times3}$. We transform the camera pose of the j-th frame to be relative to the i-th frame for simplicity, thus it holds that $E_{ij} = T_{i\to j} \times R_{i\to j}$, where $R_{i\to j} \in \mathbb{R}^{3\times3}$ and $T_{i\to j} \in \mathbb{R}^3$ are the relative rotation matrix and translation vector, respectively. Due to the contiguous representation of the epipolar line $l_{ij} = Ax + By + C$, we convert it to attention mask by calculating per-pixel distance D at coordinate (u', v') on the j-th frame to the epipolar line as

$$D_{ij}(u', v') = \frac{(A, B, C) \cdot (u', v', 1)}{\sqrt{A^2 + B^2}}, \tag{6}$$

and filtering out those values that are larger than a threshold δ. We empirically choose half of the diagonal of the feature grid size as the threshold. This approach optimizes the correspondence search space by significantly reducing the number of candidates from hw to l, with $l \ll hw$, thereby enhancing efficiency and accuracy.

Epipolar attention. We extend current temporal attention with epipolar constraint to leverage cross-frame relationship and inject geometry consistency for video generation.

We denote the query, key and value as $q \in \mathbb{R}^{hw\times c}$, $k \in \mathbb{R}^{Nhw\times c}$ and $v \in \mathbb{R}^{Nhw\times c}$, respectively. Given the epipolar attention mask $m \in \mathbb{R}^{hw\times Nhw}$ introduced in Section 3.3, our epipolar attention that captures relevant contextual information between the i-th frame and all N frames is then computed as

$$\mathrm{EpipolarAttn}(q, k, v, m) = \mathrm{softmax}\left(\frac{qk^{\mathrm{T}}}{\sqrt{d}} \odot m\right) v, \tag{7}$$

where $\odot$ denotes Hadamard product and d is the dimension of attention heads for attention score normalization. For detailed computation procedures, please refer to Appendix A.

Register tokens for scenarios where epipolar lines disappear. For videos with significant camera movements, dynamic objects, or occlusions, there may be cases where some pixels from the i-th frame have no corresponding epipolar lines within the image planes of all N frames. This situation can lead to a zero epipolar mask, affecting the computational stability of the epipolar attention mechanism.

To address this issue, we draw inspiration from Timothée et al. (2024) and introduce additional register tokens to the input sequence as a straightforward solution, as illustrated in Fig. 6. Additionally, register tokens are learnable, enabling adaptive learning to address various special cases. Without register tokens to serve as placeholders, we may encounter the zero length of key/value tokens and fail to calculate attention

3.4 Multiple Classifier-free Guidance

Control for multiple condition. Similar to DynamicCrafter (Xing et al., 2023; Esser et al., 2023), we introduce two guidance scales $s_{\text{img\&txt}}$ and s_{camera} to text-conditioned image animation, which can be adjusted to trade off the impact of two control signals:

$$\begin{aligned}\hat{\epsilon}_\theta\left(\mathbf{z}_t, \mathbf{c}_{\text{camera}}, \mathbf{c}_{\text{img\&txt}}\right) = {} & \epsilon_\theta\left(\mathbf{z}_t, \mathbf{c}_{\text{camera}}, \varnothing\right) \\ & + s_{\text{img\&txt}}\big(\epsilon_\theta\left(\mathbf{z}_t, \mathbf{c}_{\text{camera}}, \mathbf{c}_{\text{img\&txt}}\right) - \epsilon_\theta\left(\mathbf{z}_t, \mathbf{c}_{\text{camera}}, \varnothing\right)\big) \\ & + s_{\text{camera}}\big(\epsilon_\theta\left(\mathbf{z}_t, \mathbf{c}_{\text{camera}}, \mathbf{c}_{\text{img\&txt}}\right) - \epsilon_\theta\left(\mathbf{z}_t, \varnothing, \mathbf{c}_{\text{img\&txt}}\right)\big).\end{aligned} \tag{8}$$

Multiple scale distillation for acceleration. If needed, we can distill (Xing et al., 2023) the two guidance scales $s_{\text{img\&txt}}$ and s_{camera} into the model to further avoid the extra inference time brought by three times of forward:

$$\epsilon_\theta\left(\mathbf{z}_t, \mathbf{c}_{\text{camera}}, \mathbf{c}_{\text{img\&txt}}, s_{\text{camera}}, s_{\text{img\&txt}}\right) = \hat{\epsilon}_\theta\left(\mathbf{z}_t, \mathbf{c}_{\text{camera}}, \mathbf{c}_{\text{img\&txt}}\right) \tag{9}$$

4 Metrics and Evaluation

In this section, we present our reproducible evaluation pipeline. Previous studies have employed various evaluation protocols, resulting in inconsistent metrics due to the lack of a common benchmark. The structure-from-motion (SfM) method such as COLMAP (Schonberger & Frahm, 2016), struggles to produce stable and accurate predictions when applied to generated videos. This challenge arises because SfM relies on SIFT operators for keypoint identification, which can lead to erroneous matches when assessing generated content. Such inaccuracies may result in unsolvable equations or significantly flawed estimates of camera extrinsics. Contributing factors include the low resolution of these videos (256x256), the presence of dynamic scenes, the absence of true 3D consistency, and issues related to lighting variations and object distortion.

To address these limitations, we adapt the global structure-from-motion method GLOMAP (Pan et al., 2024) to validate camera pose consistency. Our evaluation pipeline comprises three steps: feature extraction, exhaustive matching, and global mapping. To enhance robustness, we share GT priors for camera intrinsics (f_x, f_y, c_x, c_y) and allow the structure-from-motion process to focus primarily on optimizing camera extrinsics. Detailed CLI parameters can be found in Appendix B.

Before calculating metrics, we canonicalize the estimated camera-to-world matrices by converting each frame relative to the first frame and normalizing the scene scale using the $\mathcal{L}_2$ distance from the first camera to the furthest cameras. To account for randomness introduced by GLOMAP, we conduct five individual trials for each of the 1,000 sampled videos, averaging only those trials that are successful per sample. The final metrics, including RotError, TransError, and CamMC, are averaged on a sample-wise basis.

RotError (He et al., 2024a). We evaluate per-frame camera-to-world rotation accuracy by the relative angles between ground truth rotations R_i and estimated rotations $\tilde{R}_i$ of generated frames. We report accumulated rotation error along 16 frames in radians.

$$\text{RotErr} = \sum_{i=1}^{n} \cos^{-1} \frac{\text{tr}(\tilde{R}_i R_i^{\text{T}}) - 1}{2} \tag{10}$$

Table 1: **Quantitative comparison with state-of-the-art methods.** * denotes the results we reproduced using DynamiCrafter as base I2V model. We achieve a 32.96%, 25.64%, 20.77% improvement over previous Sota CameraCtrl on RotErr, CamMC, TransErr on the RealEstate10K dataset without compromising dynamics, generation quality, and generalization on out-of-domain images. These results were obtained using Text and Image CFG set to 7.5, 25 steps, and camera CFG set to 1.0 (no camera cfg).

Method	Publication	TransErr ↓	RotErr ↓	CamMC ↓	FVD ↓ VideoGPT	FVD ↓ StyleGAN
DynamiCrafter (Xing et al., 2023)	ECCV 2024	9.8024	3.3415	11.625	106.02	92.196
+ MotionCtrl (Wang et al., 2024d)*	SIGGRAPH 2024	2.5068	0.8636	2.9536	70.820	60.363
+ CameraCtrl (He et al., 2024a)*	arXiv 2024	1.9379	0.7064	2.3070	66.713	57.644
+ CamI2V (Ours)		**1.4955**	**0.4758**	**1.7153**	**66.090**	**55.701**

TransError (He et al., 2024a). We evaluate per-frame camera trajectory accuracy by the camera location in the world coordinate system, i.e. the translation component of camera-to-world matrices. We report the sum of $\mathcal{L}_2$ distance between ground truth translations T_i and generated translations $\tilde{T}_i$ for all 16 frames.

$$\text{TransErr} = \sum_{i=1}^{n} \left\| \tilde{T}_i - T_i \right\|_2 \tag{11}$$

CamMC (Wang et al., 2024d). We also evaluate camera pose accuracy by directly calculating $\mathcal{L}_2$ similarity of per-frame rotations and translations as a whole. We sum up the results of 16 frames.

$$\text{CamMC} = \sum_{i=1}^{n} \left\| \left[\tilde{R}_i | \tilde{T}_i\right] - \left[R_i | T_i\right] \right\|_2 \tag{12}$$

FVD (Unterthiner et al., 2018). Additionally, to ensure that proposed method coherently improve generative capability and visual quality of base I2V model, we evaluate the distance of generated frames from training distribution by Fréchet Video Distance (FVD).

5 Experiments

5.1 Setup

Dataset. We train our model on RealEstate10K (Zhou et al., 2018) dataset, which contains approximately 70K video clips at the resolution of around 720P with camera poses annotated by SLAM-based methods. We resize video clips from dataset to 256 while keeping the original aspect ratio and perform center cropping to fit in our training scheme. We sample 16 frames from single video clip when training with a random frame stride ranging from 1 to 10. We set fixed frame stride of 8 for inference. We take random condition frame for generation as data augmentation.

Implementation Details. We choose DynamiCrafter (Xing et al., 2023) as our base image-to-video (I2V) model and implement proposed method on the top of it. For fair comparision, we also make reproduction work of MotionCtrl (Wang et al., 2024d) and CameraCtrl (He et al., 2024a), since their public accessible versions are either T2V or SVD-based. We project Plücker embedding into base model by a pose encoder similar to the architecture in CameraCtrl. We freeze all parameters from base model and train proposed method at the resolution of 256×256. We set 2 register tokens for the epipolar module to attend when no relevant pixels are on the epipolar line. We apply the Adam optimizer with a constant learning rate of 1×10^{-4}. We follow DynamiCrafter to choose Lightning as our training framework with mixed-precision fp16 and DeepSpeed ZeRO-1. We train proposed method and variants on 8 NVIDIA 3090 GPUs with effective batch size of 64 for 50K steps.

5.2 Quantitative Comparison

We compare our CamI2V with the latest methods in camera controlled image-to-video generation, including DynamiCrafter (Xing et al., 2023), MotionCtrl (Wang et al., 2024d) and CameraCtrl (He et al., 2024a). As reported in Table 1, our CamI2V significantly improves the camera controllability and visual quality, with substantial reductions in RotErr, TransErr, CamMC and FVD. Compared

Table 2: **Ablation study on model variants.** ◯ denotes our implementation of epipolar attention only on reference frame, similar to CamCo. Our proposed method (Plücker embedding along with epipolar attention on all frames) achieves SOTA performance among all variants.

Method	Plücker	Epipolar	3D Full	TransErr ↓	RotErr ↓	CamMC ↓	FVD ↓ VideoGPT	FVD ↓ StyleGAN
DynamiCrafter +CamI2V (Ours)	✓	✓		**1.4955**	**0.4758**	**1.7153**	<u>66.090</u>	**55.701**
	✓	◯		<u>1.6014</u>	<u>0.5738</u>	<u>1.8851</u>	66.439	56.778
	✓		✓	1.8215	0.6299	2.1315	71.026	60.000
	✓			1.8877	0.7098	2.2557	**66.077**	<u>55.889</u>
		✓		5.5119	1.3988	6.2855	92.605	81.447
DynamiCrafter				9.8024	3.3415	11.625	106.02	92.196

to CameraCtrl, our method reduces RotErr by 0.2306, translating to a $13.21°$ decrease in rotational error, which marks a significant improvement. And our method surpasses the state-of-the-art method CameraCtrl in other camera controllability and FVD metrics.

5.3 Ablation Study

Adding more conditions to generative models typically reduces uncertainty and improves generation quality (e.g. providing detailed text conditions through recaption). In this paper, we argue that it is also crucial to consider *noisy conditions* like latent features z_t, which contain valuable information along with random noise. For instance, in SDEdit (Meng et al., 2021) for image-to-image translation, random noise is added to the input z_0 to produce a noisy z_t. The clean component z_0 preserves overall similarity, while the introduced noise leads to uncertainty, enabling diverse and varied generations.

In this paper, we argue that **providing the model with more noisy conditions, especially at high noise levels, does not necessarily reduce more uncertainty, as the noise also introduces randomness and misleadingness**. This is the key insight we aim to convey.

To validate this point, we designed experiment with the following setups:

1. **Plücker Embedding (Baseline)**: This setup, akin to CameraCtrl, has minimal noisy conditions on cross frames due to the inefficiency of the indirect cross-frame interaction (spatial and temporal attention).
2. **Plücker Embedding + Epipolar Attention only on reference frame**: Similar to CamCo, this setup treats the reference frame as the source view, enabling the target frame to refer to it. It accesses **a small amount** of noisy conditions on the reference frame. However, some pixels of the current frame may have no epipolar line interacted with reference frame, causing it to degenerate to a CameraCtrl-like model without epipolar attention.
3. **Plücker Embedding + Epipolar Attention (Our CamI2V)**: This setup can impose epipolar constraints with all frames, including adjacent frames that have interactions in most cases to ensure an sufficient amount of noisy conditions.
4. **Plücker Embedding + 3D Full Attention**: This configuration allows the model to directly interact with features of all other frames, accessing the most noisy conditions.

The amount of accessible noisy conditions of the above four setups increase progressively. One might expect that 3D full attention, which accesses the most noisy conditions, would achieve the best performance. However, as shown in Tab. 2, 3D full attention performs only slightly better than CameraCtrl and is inferior to CamCo-like setup who only applies epipolar attention on reference frame. Notably, our method achieves best result by interacting with more noisy conditions along the epipolar lines. It can be clearly seen in the comparison part in supplementary that CamCo-like setup reference much on the first frame and cannot generate new objects. The 3D full attention generates objects within large movement due to its access to all frames pixels while it is affected by incorrect position of pixels. These findings confirm our insight that **an optimal amount of noisy conditions leads to better uncertainty reduction, rather than merely increasing the quantity of noisy conditions.**

Figure 7: Qualitative Comparison on RealEstate10K.

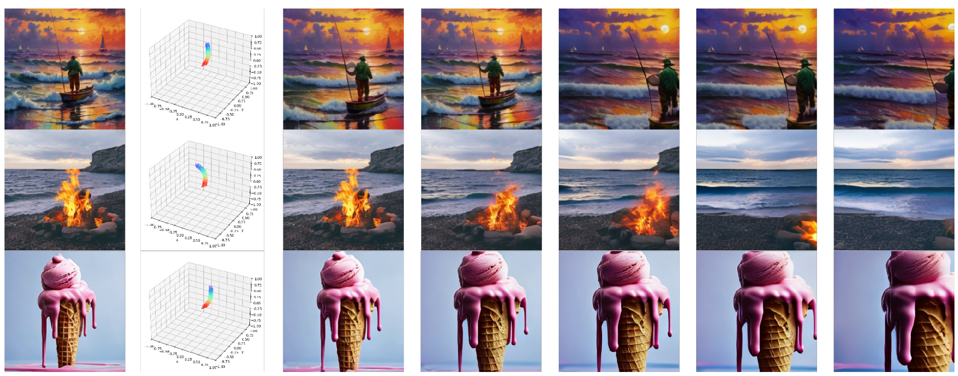
Figure 8: Out-of-Domain Visualization.

5.4 Qualitative Comparison

Visualization on RealEstate10K. As shown in Fig. 7, we present the visualization results of DynamiCrafter, MotionCtrl, CameraCtrl and our CamI2V. It can be observed that the camera trajectory of our method aligns more closely with GT compared with other methods, and the rendering of certain details appears more realistic in the video generated by our CamI2V.

Out-of-domain visualization. Our CamI2V demonstrates strong generalization capabilities, enabling direct application to camera controlled video generation across out-of-domain content, such as oil paintings, photography, and animation, as shown in Fig. 8.

6 Conclusion

In this paper, we address the integration of camera poses into diffusion models to enhance their understanding of the physical world in text-guided image-to-video generation. We propose a novel framework utilizing Plücker coordinates as 3D ray embeddings and introduce an epipolar attention mechanism that aggregates features along epipolar lines, ensuring robust tracking even under high noise conditions. Additionally, we incorporate register tokens to manage scenarios where frames lack intersections due to rapid camera movements or occlusions. Our methods significantly improve controllability and stability, achieving state-of-the-art performance on RealEstate10K and out-of-domain datasets. However, challenges remain in high-resolution generation, handling complex camera trajectories, and maintaining generation quality in long videos. Future work will focus on these aspects, alongside releasing checkpoints and training/evaluation codes to support further research.

REFERENCES

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*, 2023a.

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*, pp. 22563–22575, 2023b.

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. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

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.

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*, 2024a.

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*, 2023a.

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*, pp. 7310–7320, 2024b.

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*, 2023b.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7346–7356, 2023.

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*, pp. 22930–22941, 2023.

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.

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.

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*, 2024a.

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*, pp. 2263–2273, 2024b.

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.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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.

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*, pp. 8153–8163, 2024.

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.

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*, pp. 8079–8088, 2024.

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.

Yash Kant, Aliaksandr Siarohin, Ziyi Wu, Michael Vasilkovsky, Guocheng Qian, Jian Ren, Riza Alp Guler, Bernard Ghanem, Sergey Tulyakov, and Igor Gilitschenski. Spad: Spatially aware multi-view diffusers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10026–10038, 2024.

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*, pp. 24142–24153, 2024.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9298–9309, 2023.

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*, 2024a.

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*, volume 38, pp. 4117–4125, 2024b.

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.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14297–14306, 2023.

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*, volume 38, pp. 4296–4304, 2024.

Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes Lutz Schönberger. Global Structure-from-Motion Revisited. In *European Conference on Computer Vision (ECCV)*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

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.

Julius Plücker. *Analytisch-Geometrische Entwicklungen*, volume 2. GD Baedeker, 1828.

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.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. 2022.

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

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.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Boshi Tang, Jianan Wang, Zhiyong Wu, and Lei Zhang. Stable score distillation for high-quality 3d generation. *arXiv preprint arXiv:2312.09305*, 2023a.

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*, 2023b. URL `https://openreview.net/forum?id=2EDqbSCnmF`.

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.

Darcet Timothée, Oquab Maxime, Mairal Julien, and Bojanowski Piotr. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024.

Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, 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*, pp. 16773–16783, 2023.

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.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023a.

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*, pp. arXiv–2307, 2023b.

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, 2024a. URL https://openreview.net/forum?id=dUDwK38MVC.

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, 2024b.

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*, 2023c.

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*, 2024c.

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*, pp. 1–11, 2024d.

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*, 2024a.

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*, volume 38, pp. 6126–6134, 2024b.

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*, 2024c.

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*, 2024d.

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.

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*, 2024a.

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*, pp. 1481–1490, 2024b.

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*, pp. 1–12, 2024a.

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*, 2024b.

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.

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.

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.

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*, pp. 3836–3847, 2023a.

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*, 2023b.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7747–7756, 2024.

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*, pp. 754–769. Springer, 2022.

Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang 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*, pp. 22490–22499, 2023.

Zangwei Zheng, Xiangyu Peng, and Yang You. Open-sora: Democratizing efficient video production for all, March 2024. URL `https://github.com/hpcaitech/Open-Sora`.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *SIGGRAPH*, 2018.

A Core Codes

Algorithm 1 Spatial Attention Block

Require: U-Net feature x, condition c
1: $x \leftarrow x + \mathrm{SelfAttn}_1(\mathrm{PreNorm}(x))$
2: $x \leftarrow x + \mathrm{CrossAttn}_2(\mathrm{PreNorm}(x), c)$
3: $x \leftarrow x + \mathrm{FFN}(\mathrm{PreNorm}(x))$
4: **return** x

Algorithm 2 Temporal Attention Block with Camera Control

Require: U-Net feature x, condition c, plücker embedding p, epipolar attention mask m
1: $x \leftarrow x + \mathrm{Linear}(\mathrm{PreNorm}(x) + \mathrm{PreNorm}(p))$ ▷ Pücker Ray Embeddings
2: $x \leftarrow x + \mathrm{EpipolarAttn}(\mathrm{PreNorm}(x), m)$
3: $x \leftarrow x + \mathrm{SelfAttn}_1(\mathrm{PreNorm}(x))$
4: $x \leftarrow x + \mathrm{SelfAttn}_2(\mathrm{PreNorm}(x))$
5: $x \leftarrow x + \mathrm{FFN}(\mathrm{PreNorm}(x))$
6: **return** x

Algorithm 3 Epipolar Attention Mask

Require: Intrinsic matrices K, extrinsic matrices $[R|T]$, feature size $H \times W$, threshold δ
1: $E \leftarrow T \times R$ ▷ Essential matrices E
2: $F \leftarrow K^{-\mathrm{T}} \cdot E \cdot K^{-1}$ ▷ Fundamental matrices F
3: $g \leftarrow \mathrm{mesh\_grid}(H, W)$ ▷ Homogeneous feature coordinates g
4: $l \leftarrow \mathrm{normalize}\big(F \cdot g^{\mathrm{T}}\big)$ ▷ Epipolar line $l = Ax + By + C$, normalized by $\sqrt{A^2 + B^2}$
5: $d \leftarrow l^{\mathrm{T}} \cdot g^{\mathrm{T}}$ ▷ Distance d from feature coordinates to epipolar lines
6: $m \leftarrow$ `[reg]` $\oplus$ flatten$(d < \delta)$ ▷ Epipolar attention mask m
7: **return** m

B Colmap & Glomap Configuration

We assume `SIMPLE_PINHOLE` as the common camera model for all video clips and all 16 frames from the same video clip share the same camera intrinsics. For the feature extractor, we enable `estimate_affine_shape` and `domain_size_pooling` in `SiftExtraction`, while fix camera intrinsics by passing (f_x, f_y, c_x, c_y) into `ImageReader.camera_params`. For the exhaustive matcher, we enable `guided_matching` and set `max_num_matches` to 65536 in `SiftMatching` to make possible more underlying matches. For the global mapper, we disable `BundleAdjustment.optimize_intrinsics` and relax the geometric constraint by extending `RelPoseEstimation.max_epipolar_error` to 4.

C GPU Memory and Speed

Table 3: **Comparison on GPU memory usage and speed under DeepSpeed ZeRO-1.** * denotes our reproduction on DynamiCrafter. We report full parameter fine-tuning results of DynamiCrafter. Our model can be trained on 24GB consumer-level GPUs despite the additional epipolar attention.

Method	# Params Trainable	GPU Memory (GiB) ↓		Time (s) ↓		
		Inference	Training	Forward	Backward	Optimizer
DynamiCrafter	1.4 B	11.14	23.72	0.413	0.856	1.959
DynamiCrafter+MotionCtrl*	63.4 M	11.18	16.75	0.387	0.198	0.636
DynamiCrafter+CameraCtrl*	211 M	11.56	18.44	0.398	0.247	0.723
DynamiCrafter+CamI2V (Ours)	261 M	11.67	21.71	0.403	0.458	0.974

D Extra Out-of-Domain Visualizations

Dynamic videos are best viewed at our **local anonymous web page**. It's strongly recommended to view the visualizations in the supplementary for a more comprehensive evaluation.

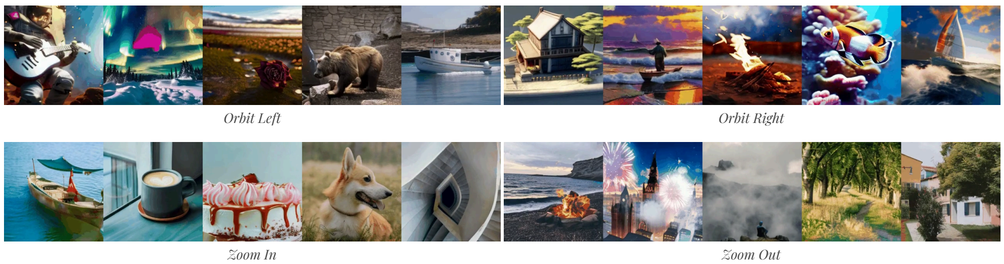

Figure 9: Visualization of our 256×256 model.

Figure 10: Visualization of original outputs from our 512×320 model, with no padding removed.

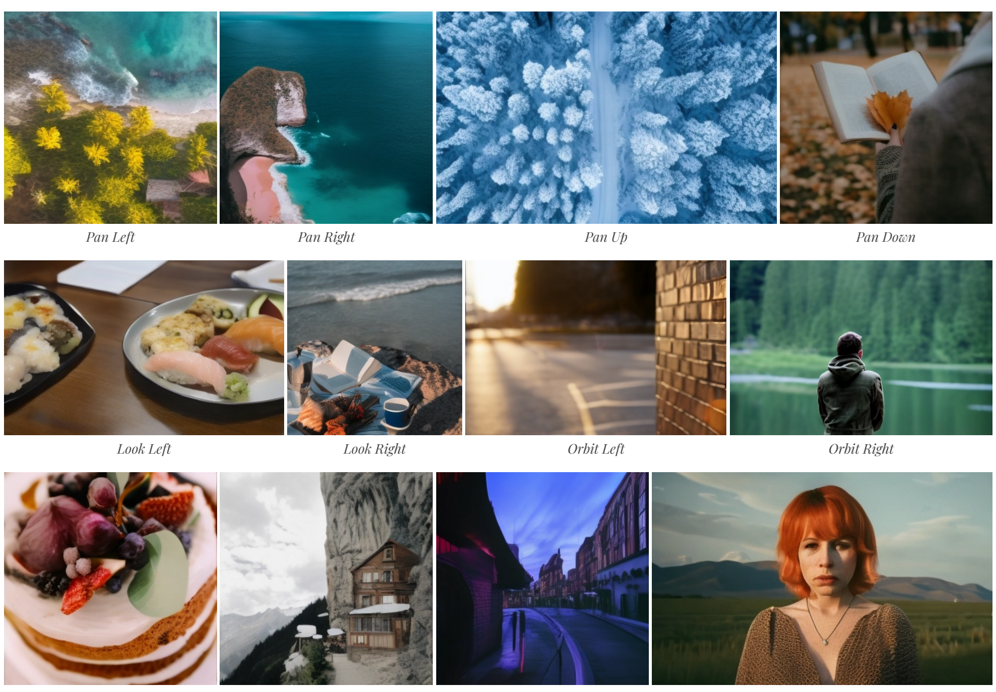

Figure 11: Generated by our 512×320 model, compatible with input images of arbitary aspect ratio.