UniScene-MoTion: Unified Scene & Motion-aware Diffusion Transition Framework

Rui Jiang[1]*, Chongmian Wang[1]*, Xinghe Fu[1], Yehao Lu[1], Teng Li[1], Xi Li[1]†

[1]College of Computer Science and Technology, Zhejiang University
{jrss, chongmianwang, xinghefu, luyehao, tengli19, xilizju}@zju.eud.cn

Abstract

Video transitions are critical for ensuring temporal coherence in edited media, yet existing methods often rely on handcrafted effects or relative-scale trajectories that fail to capture the physical structure of real-world scenes. In this work, we introduce a scale-aware video transition framework that explicitly incorporates depth-aware 3D reasoning into a diffusion-based generation pipeline. Built upon a powerful I2V foundation, our method leverages single-image depth prediction to align camera motion with metric-scale geometry, enabling physically consistent transitions. To reduce reliance on precise camera inputs, we propose a bidirectional conditional control module and a progressive training strategy with conditional dropout, enhancing generalization to loosely specified or missing camera trajectories. Extensive experiments demonstrate that our approach achieves state-of-the-art performance, delivering realistic, geometrically coherent transitions across diverse scenes and applications with minimal input guidance.

Introduction

Video transitions are a critical component in modern video editing, serving to ensure temporal continuity and narrative coherence across clips. Traditional approaches—such as hard cuts, dissolves, and wipes—offer simplicity and efficiency but often fail to capture the underlying geometry and motion dynamics of complex scenes, resulting in perceptual discontinuities and unnatural visual flow (Wolberg 1998; Shechtman et al. 2010). Recent deep learning-based (Chen et al. 2023b; Wan et al. 2025; Zhu et al. 2025; Danier, Zhang, and Bull 2024; Jain et al. 2024) approaches have sought to overcome these limitations by introducing semantically aware auxiliary constraints during generation, enabling more structured and controllable spatio-temporal interpolations. However, achieving realistic transitions in real-world scenarios remains challenging, particularly when camera motion is involved, as it requires accurate understanding of scene geometry, 3D object positions, and dynamic motion to avoid spatial misalignment.

With recent progress in image-to-video generation, the controllability of video synthesis has been greatly enhanced, primarily through two paradigms: text-driven and camera-driven approaches. For instance, a series of Diffusion-model-based works, such as I2VGen (Zhang et al. 2023), EasyAnimate (Xu et al. 2024), and AnimateAnything (Lei et al. 2025), have demonstrated the ability to achieve coarse-grained control over camera motion in generated videos via text descriptions. While these text-driven camera control methods (Zhang et al. 2023; Tian et al. 2025; **?**; Ma et al. 2025; Wan et al. 2025) are intuitive and easy to use, their expressive power often falls short in scenarios demanding precise control over camera parameters (e.g., camera position, scale, and complex motion trajectories). To mitigate this, camera-trajectory-guided models (Xu et al. 2024; Lei et al. 2025) introduce user-defined motion paths, enhancing accuracy and temporal alignment.

While trajectory-guided methods have shown promising results in video generation, they still suffer from two key limitations from observation. First, most of these methods rely on relative-scale camera trajectories, which are not grounded in the real-world metric scale. As a result, although they can capture structural relationships between objects, they lack awareness of the overall physical scale of the scene. This becomes especially problematic in tasks like video transition generation, where accurate depth perception is crucial. Without proper scale, the camera motion may produce unrealistic effects—for example, objects may appear to change size unnaturally, or foreground and background elements may shift in confusing ways—leading to visually inconsistent and disorienting results. Second, there is an ongoing challenge in balancing generation quality with usability. While precise camera trajectories can guide realistic motion, they are often hard to obtain and require expert input. On the other hand, text-guided methods are much more user-friendly, but may lack the same level of control and realism. Thus, a key question remains: how can we design models that maintain high-quality generation while being as accessible and easy to guide as language-driven approaches, without heavily relying on accurate camera inputs?

To achieve the real-world video transition effect, we intend to explicitly bring the scale prior into the video transition pipeline in this work. As depth conveys absolute scale and complements camera trajectories, we posit that incorporating estimated scene depth as an explicit prior enables physically consistent camera motion aligned with true scene

*These authors contributed equally.
†Corresponding author.

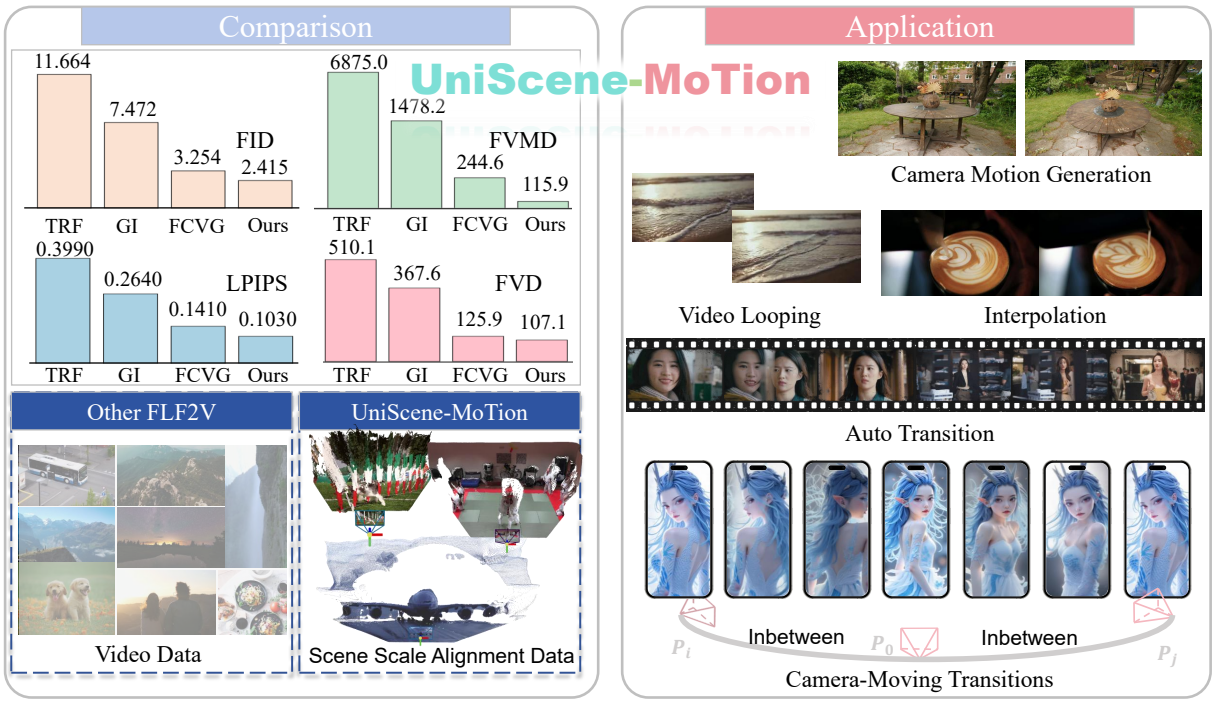

Figure 1: The comparison and application of UniScene-MoTion. Compared with previous work with raw video data, the proposed UniScene-MoTion emphasize the importance of the 3D real-scene scale information and introduce the scale prior to guide the training of the controller network. UniScene-MoTion achieves state-of-the-art generation quality in the transition task and shows potential in different applications like video looping, interpolation, auto transition, and so on.

geometry. Specifically, depth information enables us to establish precise pixel-wise correspondences between different input frames, thereby guiding the Diffusion model to generate camera motions and scene transformations that conform to physical conditions and spatio-temporal consistency. More importantly, the introduction of the prior provides users with a more structured, interpretable, and controllable interaction space: they can plan camera paths based on an understanding of scene depth, achieving fine-grained control over the transition effects as shown in Fig. 1.

In this work, we propose a scale-aware video transition framework that can be effectively controlled without relying on highly accurate camera input. The framework integrates metric-scale trajectory reasoning into a Diffusion-based video generation pipeline, leveraging single-image depth prediction. This reasoning process allows camera motion to serve as a physically meaningful prior, injecting real-world scale awareness during training. To alleviate the reliance on precise input conditions, we design a conditional control network equipped with bidirectional encoder blocks, which allow effective information exchange between the condition encoder and the frozen generation network. Furthermore, we adopt a progressive multi-stage training strategy, where the model gradually increases the output resolution while applying partial masking to the camera condition input. This reduces dependence on exact camera trajectories and improves generalization to imprecise inputs. The contributions of this work are as follows:

- We propose a scale-aware video transition framework that incorporates camera motion and depth information aligned under a consistent metric scale, providing physically grounded guidance for generation.
- We design a conditional control module with bidirectional interaction and a progressive training strategy that improves resolution and robustness while reducing reliance on precise camera inputs.
- We conduct comprehensive quantitative and qualitative experiments, demonstrating that our method achieves state-of-the-art performance on challenging video transition applications.

Related Work

Video Diffusion Models Notable works such as Sora (Brooks et al. 2024), Align Your Latents (Blattmann et al. 2023), PYoCo (Ge et al. 2023), VideoCrafter (Chen et al. 2023a, 2024), and ModelScopeT2V (Wang et al. 2023) have demonstrated exceptional performance in generating high-quality videos. Hunyuan (Team 2025) introduces a systematic framework for large video generation models, leveraging a 3D VAE and a pre-trained Multimodal Large Language Model (MLLM) to compress pixel-space videos into a compact latent space, capturing complex interactions between visual and semantic information. Wan (WanTeam 2025), a multimodal pre-trained model, excels in generating high-quality videos by processing complex text, image, and video inputs. However, these methods often rely on text or starting

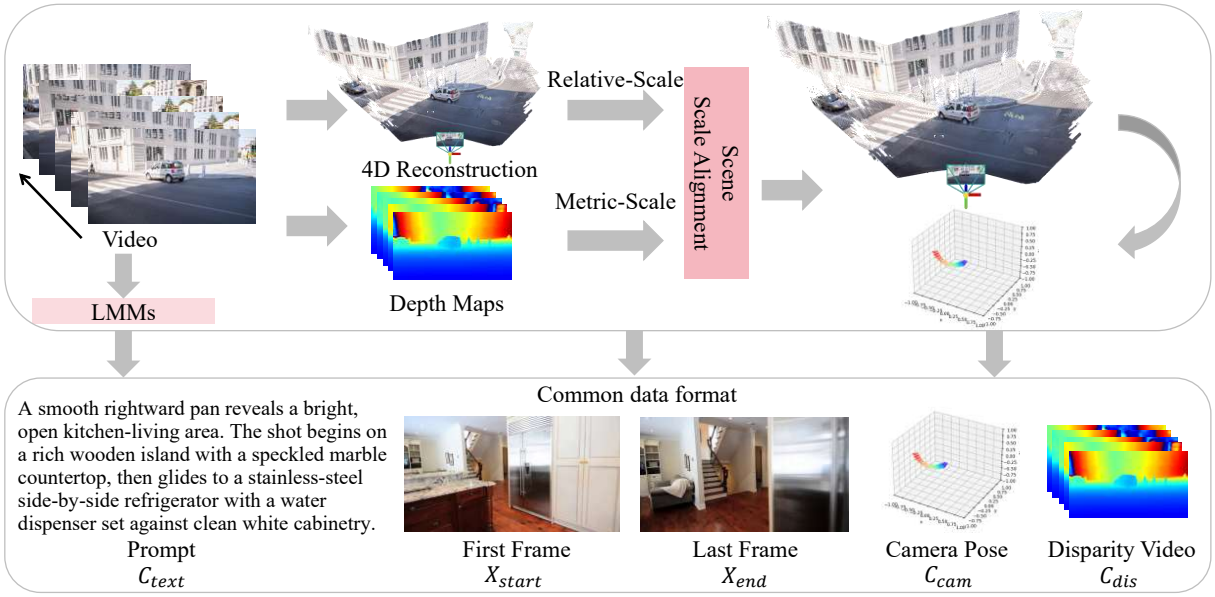

Figure 2: We illustrate the diverse data sources and key components used in our training pipeline. Multimodal large models are employed to describe camera motion directions, foreground objects, and background context. To enhance geometric stability and detail sensitivity, we convert per-frame depth maps into disparity videos, which serve as robust structural cues for 3D-aware transition generation.

image controls, which may lack the precision and interactivity required for more complex applications. To address these limitations, researchers have drawn inspiration from controllable image generation techniques like ControlNet (Zhang, Rao, and Agrawala 2023) and T2I-Adapter (Mou et al. 2024b). Several recent works have explored adding additional controls (Li et al. 2025) to video diffusion models. More recently, the focus has shifted towards motion control. Trajectory control for object motion has been introduced in works like Drag Anything (Wu et al. 2024), ReVideo (Mou et al. 2024a), and DragNUWA (Yin et al. 2023a), while camera pose control for camera motion has been explored in MotionCtrl (Wang et al. 2024b), CameraCtrl (He et al. 2024), and VD3D (Bahmani et al. 2024). In addition to these advancements, other works (Wang et al. 2025; Hong et al. 2022) have focused on enhancing the creative potential and flexibility of video generation.

Video Frame Interpolation Traditional video frame interpolation (VFI) methods, which assume moderate motion between frames, can be categorized into flow-based (Jiang et al. 2018; Xu et al. 2019; Liu et al. 2020; Niklaus and Liu 2020, 2018; Sim, Oh, and Kim 2021; Huang et al. 2020; Park, Lee, and Kim 2021) and kernel-based methods (Lee et al. 2020; Cheng and Chen 2022; Ding et al. 2021; Niklaus, Mai, and Liu 2017; Cheng and Chen 2020; Gui et al. 2020). Flow-based methods rely on optical flow estimation, while kernel-based methods use spatially adaptive kernels. However, flow-based methods suffer from inaccurate flow estimation, and kernel-based methods are limited by kernel size. Some hybrid methods combine both approaches (Bao et al. 2021; Danier, Zhang, and Bull 2022; Li et al. 2022) to address these limitations.

Several studies have explored their effectiveness for video frame interpolation, particularly in handling complex motions that are challenging for traditional optical flow-based methods. Inspired by large-scale pre-trained video diffusion models, new methods have approached VFI from a generative perspective (Danier, Zhang, and Bull 2024; Feng et al. 2024; Jain et al. 2024; Xing et al. 2023; Wang et al. 2024a). Some approaches treat input frames as conditions and train diffusion models with large-scale data, while others leverage pre-trained image-to-video diffusion models with novel sampling strategies. For instance, LDMVFI (Danier, Zhang, and Bull 2024) formulates VFI as a conditional generation problem using latent diffusion models, and VIDIM (Jain et al. 2024) employs cascaded diffusion models for high-fidelity interpolation. TRF (Feng et al. 2024) introduces a time reversal sampling strategy to fuse bidirectional motion, and Generative Inbetweening (Wang et al. 2024a) and VIBIDSampler (Yang, Kwon, and Ye 2024) further refine this approach by incorporating temporal attention and bidirectional sampling. Despite progress, these methods still struggle with large differences between starting and ending frames and typically generate a single deterministic solution without controllability.

Methods

Metric-Scale Data Alignment

High-quality video transitions rely on accurate 3D scene understanding to ensure geometric consistency and physically plausible camera motion. Without precise perception of depth, scale, and parallax, models risk producing distor-

tions and artifacts that undermine visual realism.

To support accurate 3D reasoning, we introduce metric-scale alignment of camera trajectories as a means of grounding the model's understanding of spatial structure. While scale alignment is not the end goal, it serves as an important geometric prior that helps the model interpret motion in a physically meaningful way, thereby enhancing its ability to generate coherent transitions across viewpoints. Given a video sequence $\{V_i\}_{i=1}^N$, we use per-frame depth predictions as reference to calibrate the relative-scale camera trajectories. Specifically, we convert the predicted depth maps into disparity maps, defined as the inverse of depth, to stabilize numerical behavior in far-field regions. Unlike raw depth values, which can become excessively large and noisy at long distances, disparity values are compressed and bounded, making them more robust and sensitive to near-field geometry which plays a disproportionately important role in perceived motion and transition quality.

To reinforce 3D perception and ensure geometric coherence across frames, we align the relative-scale camera trajectories—typically obtained from structure-from-motion (SfM)—with a consistent metric scale. For each frame in the video sequence, we compute two types of disparity estimates: 1) Metric-scale disparity $\{D_i^{\text{abs}}\}_{i=1}^N$, predicted using a monocular metric depth estimator (e.g., Metric3D (Yin et al. 2023b)); 2) Relative-scale disparity $\{D_i^{\text{rel}}\}_{i=1}^N$, derived from SfM reconstructions (e.g., COLMAP(Schonberger and Frahm 2016)). To align these representations, we solve for a global scene-level scale factor γ that minimizes the discrepancy between metric and relative disparities:

$$\gamma = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{N} \left\| D_i^{\text{abs}} - \gamma \cdot D_i^{\text{rel}} \right\|_2^2$$

To ensure stable scale alignment, we discard unreliable disparity values at the upper and lower 5% extremes corresponding to noisy near- and far-depth regions and retain only pixels with confidence scores in the top 50th percentile. The optimal scale factor γ^* can then be estimated via closed-form least squares:

$$\gamma^* = \frac{\sum_{i=1}^{N} D_i^{\text{abs}} \cdot D_i^{\text{rel}}}{\sum_{i=1}^{N} \left(D_i^{\text{rel}}\right)^2}$$

We apply γ^* to the translation vector $\mathbf{t}$ in each relative-scale extrinsic matrix to obtain the calibrated metric-scale camera pose:

$$\mathbf{E} = [\mathbf{R}, \gamma^* \cdot \mathbf{t}] \in \mathcal{R}^{3\times 4}$$

where $\mathbf{R} \in \mathcal{R}^{3\times 3}$ is the rotation matrix. This calibration step ensures that all camera poses are geometrically aligned in a physically meaningful scale, enabling consistent spatial reasoning across heterogeneous datasets. Integrating camera motion into a unified metric space provides more accurate 3D priors, enhancing transition quality with better spatial continuity and motion realism.

Prior-Guided Bidirectional Controller Network

Inspired by ControlNetXS (Zavadski, Feiden, and Rother 2024), we propose a Prior-Guided Bidirectional Controller Network within the DiT architecture, which enhances control signal efficiency and leverages textual prompts to significantly improve generation quality.

Specifically, as shown in Fig. 3, we decouple text tokens from interfering with control signals in the dedicated control layers. This is achieved by removing the influence of textual embeddings in these layers. Let $\mathbf{F}_{control}$ represent the input features of each layer on the control path, and $\mathbf{T}_{text}$ denote the textual embeddings. The textual embedding $\mathbf{T}_{text}$ is explicitly excluded from direct input to these control-specific modules. Our approach ensures that within the control path, the feature transformation is primarily driven by the control signal, rather than a combination with text.

Furthermore, a bidirectional flow of information is established. While the text influence is minimized in the control layers, the output of each control layer $\mathbf{F}'_{control}$ is then injected into the main DiT decoding path as the input of each DiT layer, interacting with the hidden states $\mathbf{H}$ from the DiT decoding path within each control layer. We utilize a Zero Up/Down Proj module to fuse the input control features $\mathbf{F}_{control}$ and the hidden states $\mathbf{H}$ at the start and the end of the control layer, mapping $\mathbf{F}_{control}$/$\mathbf{H}$ to the same dimension. The parameters of this module are zero-initialized for training. The control layer $\mathcal{G}$ can be conceptualized as follows (with the timestep embedding $\mathbf{T}_{\mathbf{timestep}}$ in the Diffusion model):

$$\mathbf{F}'_{control} = \mathcal{G}(\mathbf{F}_{control}, \mathbf{H}, \mathbf{T}_{timestep}).$$

This bidirectional interaction ensures that the control signals provide precise guidance, while the textual prompts contribute to the overall content and style of the generated output. This decoupling not only improves the stability of unconditional motion generation but also reduces the overall network complexity, resulting in a more lightweight control network with fewer trainable parameters.

Progressive Training with Dynamic Dropout

We adopt a resolution-progressive training pipeline composed of three stages, designed to gradually build 3D scene understanding, temporal consistency, and robustness to incomplete conditions. During training, we sample interpolation paths between clean data $\mathbf{x}_0 \sim p_{\text{data}}$ and noise $\mathbf{x}_1 \sim p_{\text{noise}}$, and optimize the model to match the true time derivative of the interpolated sample $\mathbf{x}(t) = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$:

$$\mathcal{L}_{\text{FM}} = \mathcal{E}_{\mathbf{x}_0, \mathbf{x}_1, t\sim\mathcal{U}(0,1)} \left[\left\| \frac{d\mathbf{x}(t)}{dt} - \mathbf{v}_\theta(\mathbf{x}(t), t, \mathcal{C}) \right\|^2 \right]$$

In the first stage, we train on low-resolution inputs (e.g., $81 \times 224 \times 448$) using the start frame, camera trajectory, and text prompt. This configuration encourages the model to learn plausible camera motion and scene dynamics from a single visual anchor, guided by semantic context and explicit 3D motion cues. To further enhance the model's understanding of scene geometry, we introduce an auxiliary supervision signal in the form of a ground-truth disparity video. Specifically, we take the final output feature $\mathbf{f} \in \mathcal{R}^{T\times H\times W\times C}$ from the control network and project it through a linear head $\mathcal{H}_{\text{disp}}$ to obtain predicted disparity maps $\hat{\mathbf{D}} \in \mathcal{R}^{T\times H\times W}$.

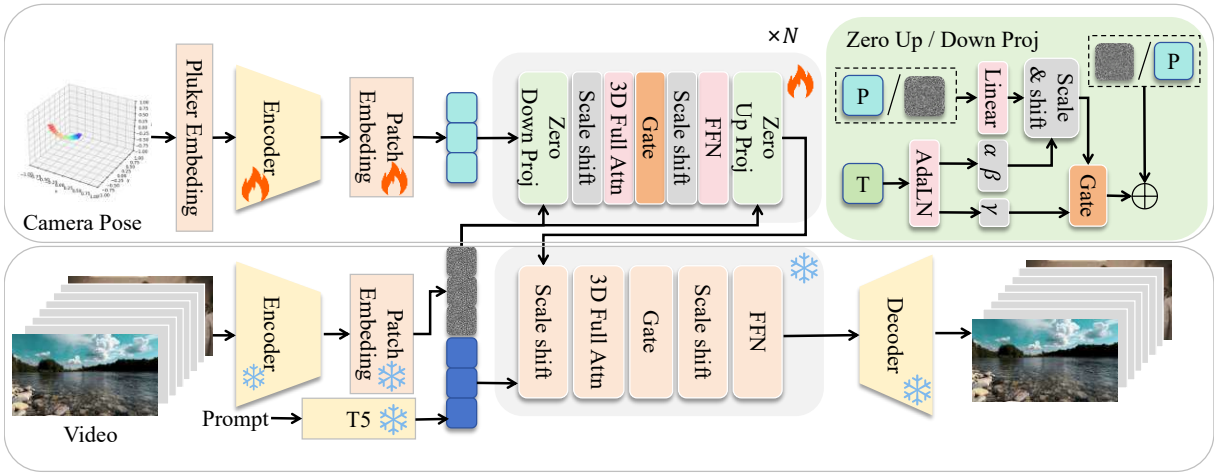

Figure 3: Pipeline of UniScene-MoTion. A lightweight plug-and-play scale-aware controller network is trained to leverage the scale prior in the generation, while the diffusion pipeline is kept frozen during training. The scale-aligned camera pose is taken as the control signal and encoded as embeddings. The Zero Up/Down Proj module is designed as a fusion module and used in each block of the controller network. This module fused the hidden states from the DiT encoder, which efficiently combine the precise guidance with the global content and style context and improves the stability of unconditional motion generation.

We supervise this predicted disparity sequence using a mean squared error loss against the annotated disparity video $\mathbf{D}^{\text{GT}}$:

$$\mathcal{L}_{\text{disp}} = \frac{1}{THW} \sum_{t=1}^{T} \left\| \hat{\mathbf{D}}_t - \mathbf{D}_t^{\text{GT}} \right\|_2^2$$

The goal is to strengthen the model's 3D motion reasoning and conditional generation from minimal input.

The second stage introduces the end frame $\mathbf{x}_{\text{end}}$, providing stronger temporal constraints for learning plausible in-between dynamics. This setup encourages the model to perform content-preserving interpolation across space and time, improving alignment and object consistency between the initial and final states.

In the final stage, we increase input resolution (e.g., $81 \times 768 \times 1360$) and apply random conditional dropout to encourage flexible generation under weak or missing conditions. While text prompts remain present throughout, we randomly drop the camera trajectory with preset probabilities:

$$\hat{\mathbf{c}}_i = \begin{cases} \mathbf{c}_i, & \text{with probability } 1 - p_i \\ \mathbf{0}, & \text{with probability } p_i \end{cases} \quad \text{for } \mathbf{c}_i \in \{\mathbf{c}_{\text{cam}}\}$$

The dropout mechanism improves robustness to incomplete control, enabling smooth and geometrically consistent transitions even with only coarse or no trajectory input—aligning with real-world usage scenarios.

Experiments

Experimental Setup

Datasets. To comprehensively evaluate the robustness and generalization ability of our method across diverse scenarios, we construct two evaluation datasets. The first dataset consists of 500 video clips selected from RealEstate10K (Zhou et al. 2018), covering a wide range of real-world indoor and outdoor environments with diverse camera motions and geometric structures. To further verify the effectiveness of our approach in varied visual and motion contexts, we additionally curate a set of 50 high-quality video clips from Pexels. This supplementary dataset spans a broad spectrum of scene types, including natural landscapes, indoor/outdoor environments, portraits, culinary scenes, and artistic styles and encompasses various motion patterns such as camera movement, object dynamics, human actions, and facial expression changes.

Evaluation metrics. Following previous works (Wang et al. 2024a; Feng et al. 2024), we adopt LPIPS (Zhang et al. 2018) and Fréchet Inception Distance (FID) (Heusel et al. 2017) to evaluate the quality of individual frames, while employing Fréchet Video Distance (FVD) (Unterthiner et al. 2019) to assess the overall quality of videos. Additionally, we take two recently proposed metrics VBench (Huang et al. 2024) and FVMD (Liu et al. 2024) to assist the evaluation, where VBench assesses videos across multiple dimensions based on pre-trained models, while FVMD refines FVD by emphasizing more on motion consistency. Furthermore, it should be noted that all these metrics are not capable of precisely evaluating temporal stability of generated videos, and thus we highly recommend directly observing more video results provided in supplementary file.

Implementation details. We adopt CogVideoX1.5-5B I2V as our base image-to-video generation model and seamlessly integrate our proposed UniScene-MoTion module as a plug-in to enhance transition quality. The entire framework is fine-tuned for 70k iterations using the AdamW optimizer with a learning rate of 1×10^{-4}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. Training is conducted on 8 NVIDIA A100 (80GB) GPUs. For each training resolution stage, we utilize the maximum batch size that fits into memory.

Figure 4: Qualitative comparison. We show the middle frame generated by each method for a given first-last frame pair to highlight the visual consistency and motion plausibility of the generation. Complete video frames are provided in the Appendix.

Method	Shape	RealEstate10K (Zhou et al. 2018)				Pexels			
		LPIPS (↓)	FID (↓)	FVMD (↓)	FVD (↓)	LPIPS (↓)	FID (↓)	FVMD (↓)	FVD (↓)
SEINE	16*768*1344	0.1213	6.206	1401.950	**106.962**	0.1339	15.180	1401.952	308.298
DynamiCrafter	16*320*512	0.2778	11.973	7865.259	268.410	0.2020	21.891	4426.780	514.185
DynamiCrafter	16*768*1344	0.2682	12.601	7905.494	283.626	0.2158	23.156	5461.399	517.670
TRF	25*768*1344	0.3988	11.664	6875.028	510.110	0.3240	29.560	3348.527	897.317
GI	25*576*1024	0.2638	7.472	1478.159	367.630	0.1920	16.028	1289.855	524.649
FCVG	25*768*1344	0.1409	3.254	244.564	125.928	0.1427	9.583	736.900	368.762
UniScene-MoTion(Ours)	25*768*1344	**0.1030**	**2.415**	**115.941**	107.133	**0.1240**	**8.964**	**537.841**	**274.478**
VideoX-Fun 1.3B	81*768*1360	0.3596	6.491	1546.225	197.752	0.2720	14.227	689.416	487.801
VideoX-Fun 14B	81*768*1360	0.3392	4.777	2135.053	218.279	0.2821	14.080	950.224	447.430
UniScene-MoTion(Ours)	81*768*1360	**0.284**	**3.746**	**227.617**	**133.213**	**0.225**	**3.746**	**549.400**	**331.285**

Table 1: Quantitative comparison on different interpolation gaps. **Bold** refer to the best results. All these metrics are not capable of precisely evaluating temporal stability of generated videos, and thus we highly recommend directly observing video results.

Comparison with State-of-the-arts

We compare UniScene-MoTion with state-of-the-art transition method SEINE (Chen et al. 2023b), and diffusion-based methods including DynamiCrafter (Xing et al. 2024), TRF (Feng et al. 2024), GI (Wang et al. 2024a), FCVG (Zhu et al. 2025), Videox-Fun and FLF2v (WanTeam 2025).

Quantitative evaluation. To evaluate performance under different motion conditions, we conduct assessments with frame gaps setting to 23 and 79 using our method. As shown in Tab. 1, our method achieves the best performance among all generative approaches across most of the metrics. When compared with short transition (the frame gap as 14) results like SEINE or DynamiCrafter, our method still performs better except on the FVD metric (the second best) with longer frame gap. Both the image-quality (FID from 9.583 to 8.964) and video-quality metrics (FVD from 308.298 to 274.478) are improved on Pexels compared with the second-best methods. The results demonstrate that our method effectively improve the generation quality in video transition by introducing the scale-prior in the genetraion. Moreover, by comparing the results under different frame gaps, SEINE may work well when the gap is small, while generative methods are more suitable for large gap. Remarkably, even with a frame gap of 79, our method achieves the best overall performance, further demonstrating its robustness.

Qualitative evaluation. Fig. 4 presents visual comparisons across several challenging transition scenarios, demonstrating the superiority of our method, UniScene-MoTion, over both early and recent two-frame-based video generation models. Early approaches such as SEINE, DynamiCrafter, and TRF often suffer from ghosting or structural distortions when facing large semantic or geometric gaps between the start and end frames. While recent large-scale models alleviate visible artifacts, they still lack explicit depth or scale reasoning, which is essential for physically coherent transitions. For instance, FunInP-14B fails to respect the physical scale of the scene—in the first row example, the phoenix remains nearly static and flies along an implausible trajectory, conflicting with the expected motion in the end frame. Similarly, FLF2V prioritizes semantic con-

Method	Frames	RealEstate10K (Zhou et al. 2018)							Pexels						
		SC	BC	MS	DD	AQ	IQ	rank	SC	BC	MS	DD	AQ	IQ	rank
SEINE	16*768*1344	0.965	**0.967**	0.988	0.246	0.512	0.729	3.67	0.951	0.963	0.988	0.162	0.519	0.610	4.00
DynamiCrafter	16*768*1344	0.932	0.942	0.972	0.600	**0.521**	0.727	4.33	0.919	0.945	0.979	0.378	0.522	0.603	4.83
TRF	25*768*1344	0.960	0.959	0.990	**0.996**	0.498	0.686	3.67	0.934	0.954	0.987	**0.850**	0.496	0.519	4.667
GI	25*576*1024	0.975	0.948	0.990	0.318	0.496	0.735	3.33	0.974	0.961	0.991	0.206	0.527	0.624	2.83
FCVG	25*768*1344	0.968	0.947	0.991	0.258	0.485	0.725	4.33	0.970	0.960	0.994	0.086	0.535	**0.632**	3.00
Ours	25*768*1344	**0.978**	0.963	**0.993**	0.656	0.514	**0.737**	**1.50**	**0.975**	**0.972**	**0.995**	0.294	**0.538**	0.624	**1.50**
VideoX-Fun 1.3B	81*768*1360	0.921	0.947	0.991	0.708	0.511	0.711	2.17	0.942	0.955	0.990	0.440	0.531	0.612	3.00
VideoX-Fun 14B	81*768*1360	0.933	**0.952**	0.989	**0.746**	**0.519**	**0.732**	**1.50**	0.945	0.956	0.990	**0.500**	**0.540**	**0.634**	**1.67**
Ours	81*768*1360	**0.939**	0.949	**0.994**	0.690	**0.519**	0.729	1.67	**0.955**	**0.962**	**0.996**	0.280	0.533	0.625	**1.67**

Table 2: A higher score indicates relatively better performance for a particular dimension. The best and second results for each column are bold and underlined, respectively. VBench-I2V includes several metrics as listed below: Subject Consistency(SC), Background Consistency (BC), Motion Smoothness (MS), Dynamic Degree (D), Aesthetic Quality (AQ), Image Quality (IQ).

	SC	BC	MS	DD	AQ	IQ
w/o Bidirection	0.929	0.943	**0.995**	**0.698**	0.518	0.708
w/o Progressive	0.926	0.945	0.994	0.692	**0.523**	0.72
Full Model	**0.939**	**0.949**	0.994	0.690	0.519	**0.729**

Table 3: Ablation study on condition components.

sistency but often resorts to abrupt masking transitions in semantically distant cases (e.g., row 2), skipping intermediate motion and resulting in unnatural temporal flow. In contrast, UniScene-MoTion integrates metric-scale trajectory reasoning and depth-aware 3D perception into a diffusion framework, enabling physically plausible and visually smooth transitions even across large content or viewpoint shifts. Our model consistently maintains structural integrity and motion coherence throughout the transition, effectively bridging semantically and geometrically distant frames. Additional qualitative results are provided in the Appendix.

Ablation Study

To assess the contribution of key components in our framework, we conduct two ablation experiments: (1) Without Bidirectional Interactive Feature Injection (denoted as w/o Bidirection), and (2) Without the progressive training schedule, trained directly on first-last frame pairs without the $\mathcal{L}_{\text{disp}}$ supervision (denoted as w/o Progressive). As shown in Tab. 3, the full model achieves the best overall performance. Removing the progressive training strategy leads to notable drops across multiple VBench metrics, indicating its importance for stable optimization and generalization. Meanwhile, removing the bidirectional interaction mechanism reduces consistency in motion and structure, confirming its role in improving temporal coherence and control fidelity.

Application

As illustrated in Fig. 5, our method demonstrates strong versatility across a range of practical video editing scenarios. First, it allows coarse camera trajectory inputs to guide the overall motion direction during transitions, offering intuitive user control. Second, our framework supports seamless transitions between distinct image-to-video sequences, enabling flexible clip stitching. This proves especially valuable for creative tasks such as AI-generated short film montage or cross-scene video composition.

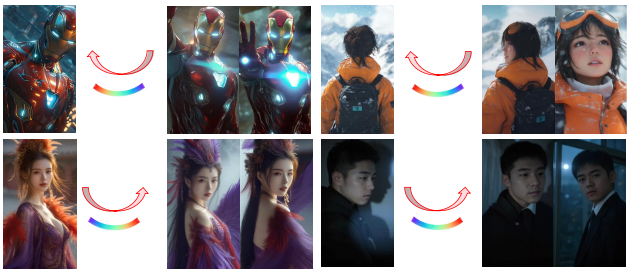

Figure 5: Applications beyond transition generation. Each row illustrates a visual generation example from our model. From left to right: (1) input first frame, (2) user-specified camera trajectory direction, (3) a representative middle frame from the generated video, and (4) the last frame.

Conclusion

We propose a depth-aware video transition framework that leverages single-frame depth estimation, disparity-guided scale calibration, and trajectory-conditioned diffusion to generate realistic and geometrically consistent camera transitions. A Bidirectional Interactive Feature Injection strategy is introduced to enhance control fidelity while maintaining semantic alignment. To improve generalization, we adopt a progressive training scheme with increasing resolution and partial trajectory dropout, enabling robustness to imprecise motion inputs. Experiments on RealEstate10K and Pexels demonstrate clear improvements over prior baselines in both visual quality and physical plausibility, establishing a unified and controllable framework for transition-aware video generation.

Acknowledgments

This work was supported in part by the National Science Foundation for Distinguished Young Scholars under Grant 62225605, Zhejiang Provincial Natural Science Foundation of China under Grant LD24F020016, NSFC under Project 12326608, Ningbo Science and Technology Special Projects under Grant 2025Z028, and the Fundamental Research Funds for Central Universities.

References

Bahmani, S.; Skorokhodov, I.; Siarohin, A.; Menapace, W.; Qian, G.; Vasilkovsky, M.; Lee, H.; Wang, C.; Zou, J.; Tagliasacchi, A.; Lindell, D. B.; and Tulyakov, S. 2024. VD3D: Taming Large Video Diffusion Transformers for 3D Camera Control. abs/2407.12781.

Bao, W.; Lai, W.; Zhang, X.; Gao, Z.; and Yang, M. 2021. MEMC-Net: Motion Estimation and Motion Compensation Driven Neural Network for Video Interpolation and Enhancement.

Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023. Align Your Latents: High-Resolution Video Synthesis with Latent Diffusion Models.

Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; Guo, Y.; Jing, L.; Schnurr, D.; Taylor, J.; Luhman, T.; Luhman, E.; Ng, C.; Wang, R.; and Ramesh, A. 2024. Video generation models as world simulators. *OpenAI technical reports*.

Chen, H.; Xia, M.; He, Y.; Zhang, Y.; Cun, X.; Yang, S.; Xing, J.; Liu, Y.; Chen, Q.; Wang, X.; Weng, C.; and Shan, Y. 2023a. VideoCrafter1: Open Diffusion Models for High-Quality Video Generation. abs/2310.19512.

Chen, H.; Zhang, Y.; Cun, X.; Xia, M.; Wang, X.; Weng, C.; and Shan, Y. 2024. VideoCrafter2: Overcoming Data Limitations for High-Quality Video Diffusion Models. abs/2401.09047.

Chen, X.; Wang, Y.; Zhang, L.; Zhuang, S.; Ma, X.; Yu, J.; Wang, Y.; Lin, D.; Qiao, Y.; and Liu, Z. 2023b. Seine: Short-to-long video diffusion model for generative transition and prediction. In *The Twelfth International Conference on Learning Representations*.

Cheng, X.; and Chen, Z. 2020. Video Frame Interpolation via Deformable Separable Convolution.

Cheng, X.; and Chen, Z. 2022. Multiple Video Frame Interpolation via Enhanced Deformable Separable Convolution.

Danier, D.; Zhang, F.; and Bull, D. 2022. ST-MFNet: A Spatio-Temporal Multi-Flow Network for Frame Interpolation.

Danier, D.; Zhang, F.; and Bull, D. 2024. LDMVFI: Video Frame Interpolation with Latent Diffusion Models.

Ding, T.; Liang, L.; Zhu, Z.; and Zharkov, I. 2021. CDFI: Compression-Driven Network Design for Frame Interpolation.

Feng, H.; Ding, Z.; Xia, Z.; Niklaus, S.; Abrevaya, V.; Black, M. J.; and Zhang, X. 2024. Explorative inbetweening of time and space. In *European Conference on Computer Vision*, 378–395. Springer.

Ge, S.; Nah, S.; Liu, G.; Poon, T.; Tao, A.; Catanzaro, B.; Jacobs, D.; Huang, J.; Liu, M.; and Balaji, Y. 2023. Preserve Your Own Correlation: A Noise Prior for Video Diffusion Models.

Gui, S.; Wang, C.; Chen, Q.; and Tao, D. 2020. FeatureFlow: Robust Video Interpolation via Structure-to-Texture Generation.

He, H.; Xu, Y.; Guo, Y.; Wetzstein, G.; Dai, B.; Li, H.; and Yang, C. 2024. CameraCtrl: Enabling Camera Control for Text-to-Video Generation. abs/2404.02101.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Hong, W.; Ding, M.; Zheng, W.; Liu, X.; and Tang, J. 2022. CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. *arXiv preprint arXiv:2205.15868*.

Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; et al. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21807–21818.

Huang, Z.; Zhang, T.; Heng, W.; Shi, B.; and Zhou, S. 2020. RIFE: Real-Time Intermediate Flow Estimation for Video Frame Interpolation. abs/2011.06294.

Jain, S.; Watson, D.; Tabellion, E.; Holynski, A.; Poole, B.; and Kontkanen, J. 2024. Video Interpolation with Diffusion Models. abs/2404.01203.

Jiang, H.; Sun, D.; Jampani, V.; Yang, M.; Learned-Miller, E. G.; and Kautz, J. 2018. Super SloMo: High Quality Estimation of Multiple Intermediate Frames for Video Interpolation.

Lee, H.; Kim, T.; Chung, T.; Pak, D.; Ban, Y.; and Lee, S. 2020. AdaCoF: Adaptive Collaboration of Flows for Video Frame Interpolation.

Lei, G.; Wang, C.; Zhang, R.; Wang, Y.; Li, H.; and Xu, W. 2025. Animateanything: Consistent and controllable animation for video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 27946–27956.

Li, C.; Wu, G.; Sun, Y.; Tao, X.; Tang, C.; and Tai, Y. 2022. H-VFI: Hierarchical Frame Interpolation for Videos with Large Motions. abs/2211.11309.

Li, T.; Zheng, G.; Jiang, R.; Wu, T.; Lu, Y.; Lin, Y.; Li, X.; et al. 2025. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. *arXiv preprint arXiv:2502.10059*.

Liu, J.; Qu, Y.; Yan, Q.; Zeng, X.; Wang, L.; and Liao, R. 2024. Fr\'echet Video Motion Distance: A Metric for Evaluating Motion Consistency in Videos. *arXiv preprint arXiv:2407.16124*.

Liu, Y.; Xie, L.; Li, S.; Sun, W.; Qiao, Y.; and Dong, C. 2020. Enhanced Quadratic Video Interpolation.

Ma, G.; Huang, H.; Yan, K.; Chen, L.; Duan, N.; Yin, S.; Wan, C.; Ming, R.; Song, X.; Chen, X.; et al. 2025. Step-video-t2v technical report: The practice, challenges,

and future of video foundation model. *arXiv preprint arXiv:2502.10248*.

Mou, C.; Cao, M.; Wang, X.; Zhang, Z.; Shan, Y.; and Zhang, J. 2024a. ReVideo: Remake a Video with Motion and Content Control. abs/2405.13865.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024b. T2I-Adapter: Learning Adapters to Dig Out More Controllable Ability for Text-to-Image Diffusion Models.

Niklaus, S.; and Liu, F. 2018. Context-Aware Synthesis for Video Frame Interpolation.

Niklaus, S.; and Liu, F. 2020. Softmax Splatting for Video Frame Interpolation.

Niklaus, S.; Mai, L.; and Liu, F. 2017. Video Frame Interpolation via Adaptive Separable Convolution.

Park, J.; Lee, C.; and Kim, C. 2021. Asymmetric Bilateral Motion Estimation for Video Frame Interpolation.

Schonberger, J. L.; and Frahm, J.-M. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Shechtman, E.; Rav-Acha, A.; Irani, M.; and Seitz, S. 2010. Regenerative morphing. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 615–622. IEEE.

Sim, H.; Oh, J.; and Kim, M. 2021. XVFI: eXtreme Video Frame Interpolation.

Team, H. 2025. HunyuanVideo: A Systematic Framework For Large Video Generative Models. arXiv:2412.03603.

Tian, J.; Qu, X.; Lu, Z.; Wei, W.; Liu, S.; and Cheng, Y. 2025. Extrapolating and Decoupling Image-to-Video Generation Models: Motion Modeling is Easier Than You Think. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12512–12521.

Unterthiner, T.; Van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; and Gelly, S. 2019. FVD: A new metric for video generation.

Wan, T.; Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.-W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; et al. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*.

Wang, J.; Yuan, H.; Chen, D.; Zhang, Y.; Wang, X.; and Zhang, S. 2023. ModelScope Text-to-Video Technical Report. abs/2308.06571.

Wang, X.; Zhou, B.; Curless, B.; Kemelmacher-Shlizerman, I.; Holynski, A.; and Seitz, S. M. 2024a. Generative inbetweening: Adapting image-to-video models for keyframe interpolation. *arXiv preprint arXiv:2408.15239*.

Wang, Y.; Chen, X.; Ma, X.; Zhou, S.; Huang, Z.; Wang, Y.; Yang, C.; He, Y.; Yu, J.; Yang, P.; et al. 2025. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, 133(5): 3059–3078.

Wang, Z.; Yuan, Z.; Wang, X.; Li, Y.; Chen, T.; Xia, M.; Luo, P.; and Shan, Y. 2024b. MotionCtrl: A Unified and Flexible Motion Controller for Video Generation.

WanTeam. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. arXiv:2503.20314.

Wolberg, G. 1998. Image morphing: a survey. *The visual computer*, 14(8-9): 360–372.

Wu, W.; Li, Z.; Gu, Y.; Zhao, R.; He, Y.; Zhang, D. J.; Shou, M. Z.; Li, Y.; Gao, T.; and Zhang, D. 2024. DragAnything: Motion Control for Anything using Entity Representation. abs/2403.07420.

Xing, J.; Xia, M.; Zhang, Y.; Chen, H.; Wang, X.; Wong, T.; and Shan, Y. 2023. DynamiCrafter: Animating Open-domain Images with Video Diffusion Priors. abs/2310.12190.

Xing, J.; Xia, M.; Zhang, Y.; Chen, H.; Yu, W.; Liu, H.; Liu, G.; Wang, X.; Shan, Y.; and Wong, T.-T. 2024. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, 399–417. Springer.

Xu, J.; Zou, X.; Huang, K.; Chen, Y.; Liu, B.; Cheng, M.; Shi, X.; and Huang, J. 2024. Easyanimate: A high-performance long video generation method based on transformer architecture. *arXiv preprint arXiv:2405.18991*.

Xu, X.; Si-Yao, L.; Sun, W.; Yin, Q.; and Yang, M. 2019. Quadratic Video Interpolation.

Yang, S.; Kwon, T.; and Ye, J. C. 2024. ViBiDSampler: Enhancing Video Interpolation Using Bidirectional Diffusion Sampler. *arXiv preprint arXiv:2410.05651*.

Yin, S.; Wu, C.; Liang, J.; Shi, J.; Li, H.; Ming, G.; and Duan, N. 2023a. DragNUWA: Fine-grained Control in Video Generation by Integrating Text, Image, and Trajectory. abs/2308.08089.

Yin, W.; Zhang, C.; Chen, H.; Cai, Z.; Yu, G.; Wang, K.; Chen, X.; and Shen, C. 2023b. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9043–9053.

Zavadski, D.; Feiden, J.-F.; and Rother, C. 2024. Controlnet-xs: Rethinking the control of text-to-image diffusion models as feedback-control systems. In *European Conference on Computer Vision*, 343–362. Springer.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, S.; Wang, J.; Zhang, Y.; Zhao, K.; Yuan, H.; Qin, Z.; Wang, X.; Zhao, D.; and Zhou, J. 2023. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*.

Zhou, T.; Tucker, R.; Flynn, J.; Fyffe, G.; and Snavely, N. 2018. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*.

Zhu, T.; Ren, D.; Wang, Q.; Wu, X.; and Zuo, W. 2025. Generative inbetweening through frame-wise conditions-driven video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 27968–27978.