FlexComposer: Unified Video Compositing from Images to Dynamic Footage with Flexible Trajectory Control

Songchun Zhang[1], Sitong Guo[2], Xianghao Xiong[1], Pengwei Liu[2], Yuwei Guo[3], Lvmin Zhang[4], and Anyi Rao[1]

[1] Hong Kong University of Science and Technology
[2] Zhejiang University
[3] The Chinese University of Hong Kong
[4] Stanford University

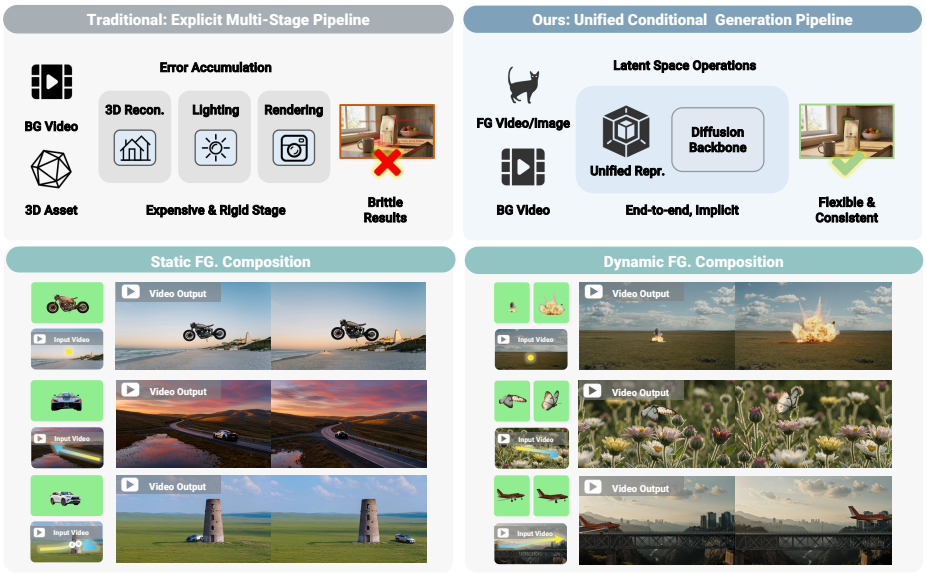

Fig. 1: Versatile Video Compositing with FlexComposer. Our framework unifies the insertion of diverse assets into target videos through user-defined trajectories (visualized as arrows in input thumbnails). **Left (Static Foreground Comp.):** Given a single static image, the model synthesizes realistic object motion and view changes while achieving photorealistic lighting adaptation (middle row) and handling depth occlusions (bottom row). **Right (Dynamic Foreground Comp.):** For video inputs, FlexComposer preserves the internal dynamics of the asset (e.g., the expansion of the explosion or the flapping of butterfly wings) while harmonizing it spatially and photometrically with the background scene.

Abstract. Generative video compositing, which involves inserting external assets seamlessly into existing video sequences, is essential for content creation and visual effects. However, existing approaches suffer from a control-fidelity trade-off: they either hallucinate motion from

static images, failing to preserve the dynamics of pre-animated assets, or lack fine-grained spatial control for precise asset placement along user-defined trajectories. We propose FlexComposer, a unified framework that standardizes video compositing as a trajectory-guided conditional generation task, enabling the seamless integration of both static images and dynamic footage. Our approach introduces three key designs: (1) a Unified Canonical Foreground Representation that decouples an object's intrinsic motion from its global displacement, standardizing heterogeneous inputs into a stabilized, centered latent space; (2) a Spatial-Aware Latent Injection strategy that exploits the translation equivariance of VAE latent spaces to transport canonical features onto target trajectories via a parameter-free mechanism; and (3) a Hybrid Dataset and Synthetic-to-Real Curriculum that synergizes procedural simulation, real-world cinematic footage, and generative data to implicitly learn physically plausible illumination and shadow harmonization. This unified design handles diverse inputs—from product photos to dynamic subjects—achieving high-fidelity motion control and environmental integration without the need for explicit 3D reconstruction or auxiliary learnable adapters. Extensive experiments demonstrate that FlexComposer outperforms state-of-the-art methods in visual quality, temporal consistency, and trajectory adherence.

Keywords: Video Diffusion Models · Trajectory Control · Generative Video Compositing

1 Introduction

The realistic insertion of virtual assets into existing video sequences is a fundamental yet challenging problem in computer vision and graphics. This capability serves as the backbone for visual effects, augmented reality, 3D-aware content creation [89, 91, 92], and the burgeoning field of personalized video content creation. Achieving seamless composites requires simultaneously preserving background integrity, maintaining spatiotemporal consistency of inserted foregrounds, and enabling precise motion control over their trajectories.

While traditional graphics pipelines [33] and harmonization techniques [22, 34] excel in rendering or color adjustment, they operate via explicit decoupled stages prone to cascading error accumulation, inherently struggling to integrate dynamic assets with coherent motion. In contrast, recent generative frameworks [10,11,31,32,37,60,61,80,82] leverage diffusion models to address these limitations. These approaches offer promising avenues for composition, ranging from zero-shot subject animators [10, 61] to unified in-context editors [11, 32, 73, 82] and large-scale foundation models [60, 80]. However, despite their versatility, a critical gap remains: most methods either hallucinate motion from static references [61] or operate at abstract semantic levels [32, 73]. Crucially, they lack a unified mechanism to seamlessly integrate both static images and dynamic video assets—preserving intrinsic fidelity while ensuring fine-grained spatial control.

To address these limitations, we propose FlexComposer, a unified conditional video compositing framework. We bridge the gap between control and fidelity through three key technical components. First, to handle heterogeneous inputs, we introduce a Unified Canonical Foreground Representation that decouples intrinsic motion from global displacement by stabilizing dynamic videos into canonical sequences at the pixel level, enabling consistent treatment of diverse inputs. Second, to achieve precise spatial control without auxiliary adapters (e.g., ControlNet [88]), we propose Spatial-Aware Latent Injection, a parameter-free mechanism that exploits VAE latent translation equivariance to directly transport canonical features onto user-defined trajectories while preserving temporal synchronization. Third, to resolve photometric inconsistencies, we construct a Hybrid Dataset combining procedural simulation data for geometric grounding, real-world footage for realism, and generative data for semantic diversity, trained through a synthetic-to-real curriculum. Extensive experiments demonstrate that FlexComposer achieves superior performance in visual quality, temporal consistency, and controllability.

In summary, our contributions are threefold: (1) we propose **FlexComposer**, a unified generative framework that seamlessly integrates both static images and dynamic video clips into target sequences by utilizing a canonical representation to decouple intrinsic object motion from global trajectory control; (2) we introduce **Spatial-Aware Latent Injection**, a parameter-free mechanism that leverages translation equivariance to enable precise spatial control while preserving the high-fidelity details of foreground assets without auxiliary adapters; and (3) we implement a **synthetic-to-real curriculum** learning strategy that leverages a diverse mixture of procedural simulation and real-world footage. This approach empowers the model to robustly generalize across diverse scenes with coherent synthesis of lighting and shadows. Extensive experiments demonstrate that FlexComposer achieves superior performance in visual quality, temporal consistency, and controllability.

2 Related Work

2.1 Video Generative Models

The landscape of video generation has shifted fundamentally from early GANs [18] and autoregressive transformers [14] to Diffusion Models (VDMs) [7, 26, 27], which now serve as the gold standard for high-fidelity generation. While Latent Diffusion Models (LDMs) [58] balanced quality and efficiency, the recent emergence of Video Diffusion Transformers (DiTs) [36, 54, 81], autoregressive video models [90], and memory-augmented long video generation [6], exemplified by Sora [9, 39, 94], has demonstrated superior scalability and temporal coherence. Despite their ability to generate photorealistic content from text, these foundation models inherently lack precise spatial controllability. Our work builds upon the powerful priors of these large-scale DiTs, bridging the gap between high-fidelity synthesis and fine-grained, user-specified control.

2.2 Motion-Controlled Video Generation

While T2V models achieve photorealism, text prompts lack the granularity required for precise control. Existing spatial guidance methods fall into three categories. First, trajectory-based control employs sparse 2D drag points or bounding boxes [21, 29, 51–53, 72, 83, 86], with recent works integrating trajectory adapters directly into DiT architectures [13, 17, 47, 66, 67, 93]. However, these methods operate in 2D screen space and fail to comprehend the underlying 3D scene geometry. To incorporate 3D geometric awareness, efforts have been directed toward camera control and 3D-consistent scene generation. Approaches range from traditional reconstruction-based methods [16, 40, 68, 77] to generative techniques using multi-view diffusion, inpainting, pose conditioning, or sparse-view reconstruction [3–5, 23, 48, 57, 59, 64, 75, 84, 85, 87, 89, 91, 92, 95]. Nevertheless, such global control focuses on the ego-motion rather than the fine-grained spatial dynamics of individual objects within the scene. Alternatively, structure-based control leverages dense priors such as skeleton, depth, or normal maps [20, 41, 49, 70]. While offering geometric guidance, these methods impose heavy data preparation burdens that limit their applicability.

2.3 Video Compositing and Harmonization

Traditional approaches to video composition employ explicit decoupled pipelines. Graphics-based methods [33] rely on 3D reconstruction, lighting estimation, and ray tracing in separate stages, with error accumulation across stages limiting robustness. Professional tools [1, 15] built on Porter and Duff's compositing operator [55] enable nondestructive layer-based workflows but require extensive manual per-frame artistry for masking and effect tuning. Recent lightweight approaches focus on color harmonization [22, 79], enabling rapid integration through color matching and tone adjustment. However, these methods operate in a rigid copy-paste paradigm without accounting for geometric interactions, lighting adaptation, or dynamic foreground motion—prerequisites for realistic asset composition. Instruction-based video editing originated from image-level methods [8] and has evolved toward video through per-video fine-tuning [56, 74], inference-only protocols [10–12, 32, 38, 50, 60, 76, 80, 82], and concept-level image/video composition [37]. While diffusion priors enable semantic editing, they lack precise trajectory mechanisms for asset insertion. Conversely, generative compositing frameworks [10, 31, 61] address insertion but primarily hallucinate motion from static references, failing to preserve the intrinsic dynamics of pre-animated assets. Our work bridges these complementary research directions by unifying foreground representation with spatial control mechanisms within a generative framework specifically designed for composition.

3 Method

Given a foreground asset V_{raw} and a background sequence V_{bg}, our goal is to synthesize a composite video where the asset follows a user-defined trajectory $\mathcal{T}_{user}$.

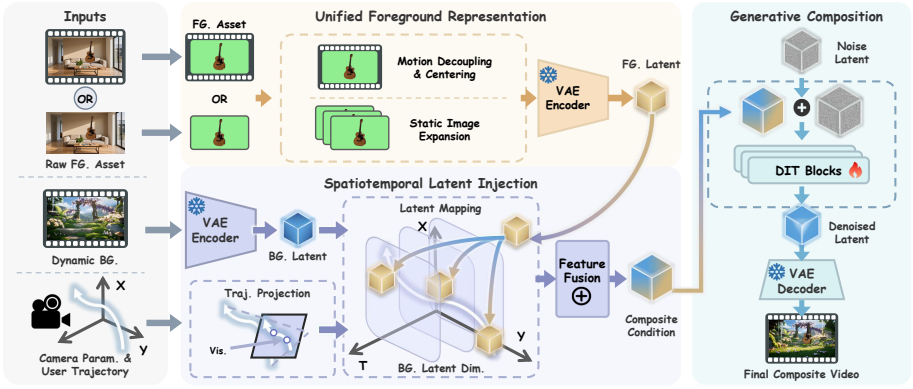

Fig. 2: The Pipeline of FlexComposer. (1) **Inputs**: A dynamic background, a static/dynamic foreground asset, a 3D trajectory, and visibility masks. (2) **Unified Representation**: Diverse inputs are encoded into a canonical feature space to preserve identity and local motion. (3) **Spatiotemporal Latent Injection**: Canonical features are warped into the background latent space along the trajectory, utilizing a visibility gate for occlusions. (4) **Generative Composition**: The warped features condition a video diffusion model to synthesize the final composite.

Users can optionally specify a visibility schedule $\mathcal{V}_{user}[t] \in \{0, 1\}$ to control when the object appears or disappears. As illustrated in Figure. 2, our framework follows a three-stage pipeline: (1) Unified Canonical Foreground Representation, which standardizes diverse inputs into a stabilized, centered latent space; (2) Spatial-Aware Latent Injection, which transports canonical features onto the target trajectory via a parameter-free mechanism; and (3) Generative Composition, which leverages a conditional diffusion transformer to synthesize the final video.

3.1 Unified Canonical Foreground Representation

We unify heterogeneous inputs (V_{raw}) into a canonical representation through pixel-space preprocessing followed by latent encoding.

Dynamic Video Stabilization. To eliminate conflicting camera ego-motion, we utilize Grounded-SAM2 [35] for segmentation and SpatialTracker v2 [78] for centroid tracking ($\mathbf{c}_t$) and visibility estimation ($\mathcal{V}_{\text{track}}$). We apply a reverse translation to align $\mathbf{c}_t$ to the canvas center, yielding a stabilized, in-place sequence $\mathbf{I}_{\text{stab}}$ where the subject animates at the center, effectively decoupling local motion from global displacement.

Static Image Expansion. For static inputs, we replicate the image to target duration T to form $\mathbf{I}_{\text{expand}}$. We further inject Gaussian noise along the temporal dimension after encoding, encouraging the generative backbone to hallucinate plausible temporal dynamics while preserving the subject's identity.

Relighting Augmentation. To prevent the network from merely copy-pasting the source appearance and to enforce photometric consistency with diverse back-

ground contexts, we apply random illumination perturbations to the frames via a relighting module [45]. The final preprocessed sequence is denoted as $\mathbf{I}_{\text{can}}$.

Latent Space Encoding. We encode the preprocessed sequence $\mathbf{I}_{\text{can}}$ using the Video VAE encoder: $\mathbf{Z}_{\text{can}} = \mathcal{E}(\mathbf{I}_{\text{can}}) \in \mathbb{R}^{T' \times H' \times W' \times C}$, where $T' = T/f_t$, $H' = H/f_s$, $W' = W/f_s$ reflect temporal and spatial downsampling factors. For static inputs, we inject Gaussian noise along the temporal axis to break frame-wise redundancy and encourage motion synthesis

3.2 Spatial-Aware Latent Injection

Unlike existing methods [17,20] that rely on auxiliary adapters which often suffer from signal degradation, we introduce a parameter-free injection strategy.

Spatiotemporal Coordinate Mapping. We first project the 3D trajectory $\mathcal{T}_{3D} = \{\mathbf{P}_k\}_{k=1}^{T}$ onto the 2D image plane to obtain pixel coordinates $\mathbf{p}_k = \Pi(\mathbf{P}_k)$, where $\Pi(\cdot)$ incorporates the camera intrinsics and extrinsics. To align these coordinates with the latent space of VAE, which undergoes spatial downsampling by factor f_s and temporal compression by factor f_t, we compute the discrete latent coordinate $\tilde{\mathbf{u}}_n$ for each latent frame index n:

$$\tilde{\mathbf{u}}_n = \left\lfloor \frac{1}{f_t} \sum_{k=(n-1)f_t+1}^{n \cdot f_t} \frac{\mathbf{p}_k}{f_s} \right\rceil, \quad 1 \le n \le \frac{T}{f_t} \tag{1}$$

where $\lfloor \cdot \rceil$ denotes the nearest integer rounding operation. This discretization ensures that the continuous projected coordinates are accurately mapped to the fixed integer grid indices of the latent feature map $\mathbf{Z}$.

Frame-wise Feature Transport. We define a transport operation Ψ that maps features from the canonical center $\mathbf{c}$ to the target trajectory $\tilde{\mathbf{u}}_n$. Let Ω denote the spatial extent (bounding box) of the foreground object in the latent space. For every local spatial offset $\boldsymbol{\delta} \in \Omega$, we map the feature at the canonical position $\mathbf{c} + \boldsymbol{\delta}$ to the target position $\tilde{\mathbf{u}}_n + \boldsymbol{\delta}$ in the composite latent $\mathbf{Z}_{\text{trans}}$:

$$\mathbf{Z}_{\text{trans}}[n, \tilde{\mathbf{u}}_n + \boldsymbol{\delta}] = \mathbf{Z}_{\text{can}}[n, \mathbf{c} + \boldsymbol{\delta}] \cdot \mathcal{V}[n, \boldsymbol{\delta}] \tag{2}$$

where $\mathcal{V}[n, \boldsymbol{\delta}] = \mathcal{V}_{\text{vis}}[n] \cdot \alpha[n, \boldsymbol{\delta}] \in \{0, 1\}$ combines a temporal visibility flag with the spatial segmentation mask $\alpha[n, \boldsymbol{\delta}]$. During training, $\mathcal{V}_{\text{vis}}$ is derived from SpatialTracker's occlusion detection; at inference, it is user-specified via $\mathcal{T}_{user}$, enabling explicit control over when the object appears along the trajectory. Crucially, by coupling the temporal index n on both sides of Eq. 2, we preserve the object's intrinsic dynamics while updating its global location.

3.3 Generative Composition

Our video synthesis leverages the Wan Video Diffusion Transformer [63] as the generative backbone. However, the standard I2V protocol is insufficient for our dynamic compositing task that requires dense spatiotemporal guidance. We therefore adapt the conditioning mechanism to accept our composite features.

Composite Feature Fusion. We fuse the transported foreground features $\mathbf{Z}_{\text{trans}}$ with the background latent $\mathbf{Z}_{\text{bg}}$ to construct the composite condition $\mathbf{C}_{\text{fuse}}$:

$$\mathbf{C}_{\text{fuse}} = \mathbf{Z}_{\text{bg}} \odot (1 - \mathbf{M}) + \mathbf{Z}_{\text{trans}} \odot \mathbf{M} \tag{3}$$

where $\mathbf{M}$ denotes the foreground mask in latent space and $\odot$ represents element-wise multiplication.

Conditional Input Formation. The network input is formed by channel-wise concatenating the noisy latent $\mathbf{x}_t$ at diffusion timestep t with the composite condition: $\mathbf{Z}_{\text{in}} = \text{Concat}(\mathbf{x}_t, \mathbf{C}_{\text{fuse}})$. This dense spatiotemporal conditioning explicitly guides generation with strong geometric priors from the transported features, unlike standard I2V methods that hallucinate motion from a single frame.

Training Objective. We optimize the model using the Flow Matching objective [44]. Crucially, the velocity prediction network $\mathbf{v}_\theta$ now takes the concatenated latent $\mathbf{Z}_{\text{in}}$ as its spatial input:

$$\mathcal{L}(\theta) = \mathbb{E}_{t,\mathbf{x}_t,\mathbf{c}_{\text{txt}}} \left[\left\| \mathbf{v}_\theta(\mathbf{Z}_{\text{in}}, t, \mathbf{c}_{\text{txt}}) - \mathbf{v}_t(\mathbf{x}_t) \right\|^2 \right] \tag{4}$$

where $\mathbf{x}_t$ denotes the flow state at timestep t, $\mathbf{v}_t$ is the ground-truth velocity field, and $\mathbf{c}_{\text{txt}}$ represents the text embeddings injected via cross-attention. This formulation ensures the diffusion process is strictly conditioned on our trajectory-aligned composite features.

3.4 Hybrid Dataset and Curriculum Learning

To achieve precise motion control and robust generalization, we introduce a hybrid dataset construction pipeline and a three-stage curriculum learning strategy. This design is crucial for progressively disentangling geometric motion from photometric appearance. Additional details are provided in the supplementary materials.

Simulation Bootstrap. We first utilize procedurally generated data from dynamic 3D characters [2] and synthetic environments [19] to establish geometric grounding. We train the model for 2k steps on 12k synthetic clips with explicit 3D vertex trajectories. This strict supervision warms up the spatial injection mechanism, enforcing the network to adhere to rigid spatial transport priors before tackling appearance harmonization.

Real-World Adaptation. To bridge the domain gap, we curate a hybrid dataset comprising 54k high-quality real-world clips, augmented with estimated 3D trajectories. In this phase, we fine-tune the model for 5k steps. Training on these canonicalized real sequences enables the model to disentangle local motion from global movement and synthesize coherent environmental interactions (e.g., shadows and lighting) that are absent in synthetic data.

Open-Domain Refinement. Finally, to support open-domain content creation, we introduce a generative synthesis pipeline prompting foundation T2I/I2V

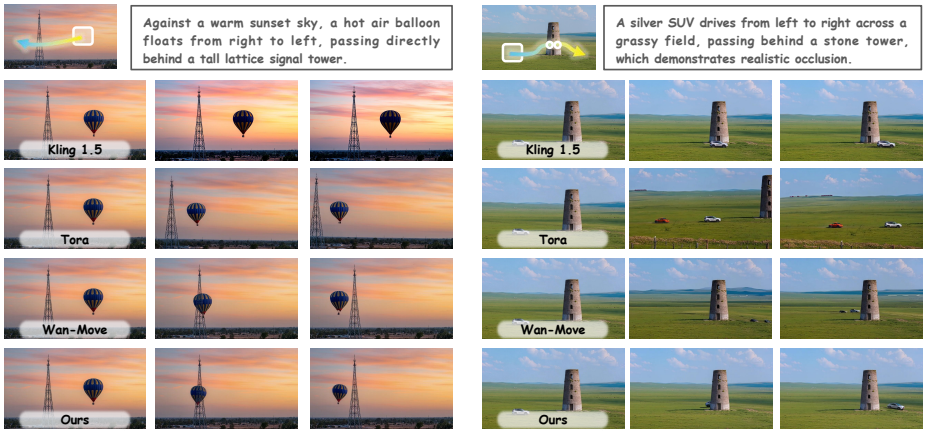

Fig. 3: Qualitative Comparison on Static Foreground Compositing. We compare our method against trajectory-controlled I2V methods on two challenging scenarios. Our method demonstrates superior occlusion handling and trajectory adherence while maintaining visual fidelity

models. We refine the model for 1k steps on 5k highly diverse clips with pseudo-ground-truth trajectories. This strategy effectively prevents overfitting to limited real-world categories and broadens semantic coverage for novel assets.

4 Experiments

To evaluate our method, we consider its nature as both a trajectory control framework and a video compositing tool. Since standard benchmarks typically focus on isolated tasks, we adopt a structured evaluation protocol across three domains: (1) *Static Foreground Compositing* against motion-controllable I2V methods; (2) *Dynamic Foreground Compositing* against video editing baselines; and (3) *Video Harmonization* against specialized harmonization algorithms.

4.1 Implementation Details

We implement FlexComposer building upon the Wan2.1-I2V-14B architecture [63], initializing the backbone with Wan-Move weights [13] to leverage robust motion priors. Unlike methods relying on heavy auxiliary networks, our spatial injection uses a parameter-free mechanism to preserve feature fidelity. To adapt the backbone without catastrophic forgetting, we employ Low-Rank Adaptation (LoRA) [28] ($r = 64$) on the projection layers of all DiT blocks. Training is conducted on 32 NVIDIA A100 GPUs using FSDP and Ulysses sequence parallelism (degree 4), optimized via AdamW ($lr = 1 \times 10^{-5}$) with a total batch size of 32. More details can be found in supplemental materials.

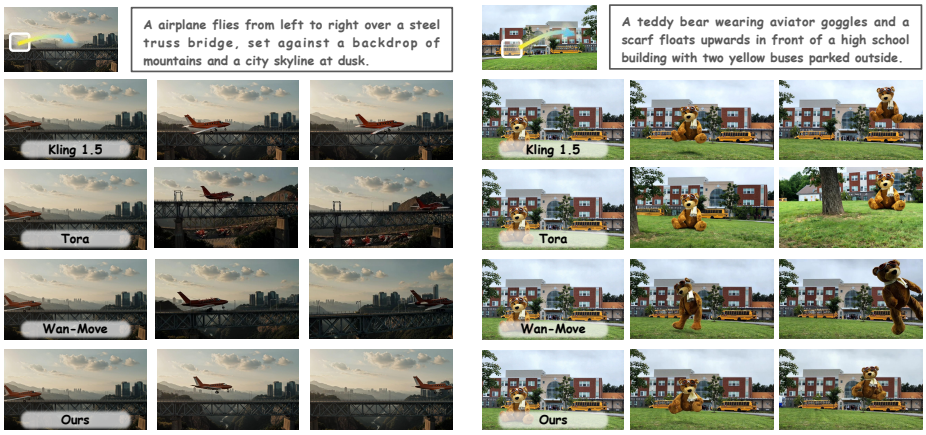

Fig. 4: Additional Qualitative Comparison on Static Foreground Compositing. We compare our method against trajectory-controlled I2V methods on two challenging scenarios.

Table 1: Quantitative comparison on DAVIS. Comparison of video quality (FID, FVD) and trajectory consistency (EPE) against recent baselines.

Method	EPE ↓	FID ↓	FVD ↓	PSNR ↑	SSIM ↑
ImageConductor [43]	14.92	54.5	512.8	11.55	0.468
LeviTor [65]	3.65	22.3	116.1	13.20	0.515
Tora [93]	3.52	26.1	128.5	13.65	0.492
MagicMotion [42]	3.48	24.5	112.9	12.90	0.535
Wan-Move [13]	<u>2.50</u>	<u>14.7</u>	94.3	16.50	0.610
Ours (I2V)	2.38	14.1	<u>89.5</u>	<u>16.85</u>	<u>0.625</u>
Ours (V2V)	**2.15**	**13.4**	**82.6**	**17.40**	**0.658**

4.2 Static Foreground Compositing

In this setting, we aim to animate a static image along a user-defined trajectory while maintaining identity.

Setup. We evaluate on DAVIS [69] and MoveBench [13]. Following standard protocols, we evaluate the fidelity of generated videos using FVD [62] and FID [25] for perceptual quality, alongside PSNR and SSIM [71] for frame-level accuracy. We compare against state-of-the-art motion-controllable methods [13, 42, 43, 65, 93] under two settings: *I2V* (first frame only) and *V2V* (full background context).

Results and Analysis. Table 1 summarizes the quantitative evaluation. In the I2V setting, our method demonstrates reliable trajectory adherence, comparing favorably to representative baselines such as ImageConductor [43] and Tora [93]. These results suggest that our trajectory injection mechanism effectively guides motion generation, offering a robust alternative to pixel-level

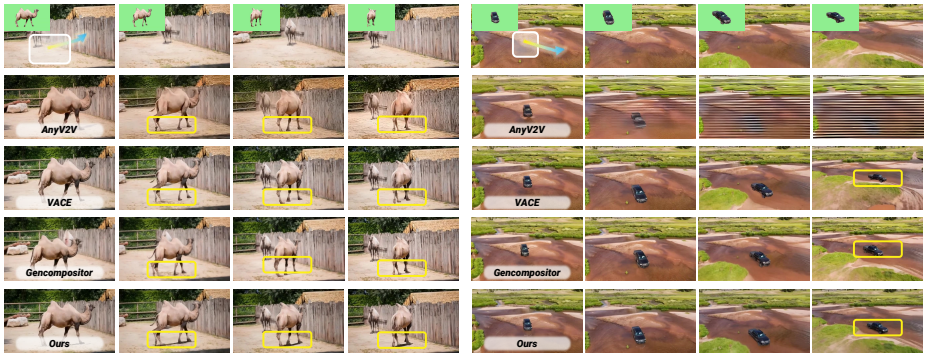

Fig. 5: Qualitative Comparison on Dynamic Foreground Compositing. We evaluate our method against video editing and compositing approaches on the V2V task.

Table 2: Quantitative comparison on the MoveBench dataset. We report trajectory consistency and video quality metrics. Bold indicates the best results, and underline denotes the second best.

Method	EPE ↓	FID ↓	FVD ↓	PSNR ↑	SSIM ↑
ImageConductor [43]	15.55	34.8	422.5	13.50	0.485
LeviTor [65]	3.45	18.3	99.2	15.55	0.538
MagicMotion [42]	3.25	17.6	97.1	14.85	0.562
Tora [93]	3.32	22.8	101.5	15.65	0.548
Wan-Move [13]	2.60	12.2	83.5	<u>17.80</u>	<u>0.640</u>
Ours (I2V)	<u>2.42</u>	<u>11.9</u>	<u>80.4</u>	17.95	0.655
Ours (V2V)	**2.38**	**11.2**	**74.9**	**18.60**	**0.682**

injection approaches for handling complex paths. Regarding visual quality, our method exhibits improved FVD and consistency. This indicates that full video context ensures temporal consistency, while our I2V variant maintains high fidelity without temporal priors.

4.3 Dynamic Foreground Compositing

This task involves transforming a dynamic video foreground. The challenge lies in altering the global trajectory while preserving the object's intrinsic dynamics. **Setup.** We evaluate on a curated test set containing 50 video assets with diverse intrinsic motions and pre-defined trajectories. We compare against Video Editing & Compositing methods: AnyV2V [38], VACE [31], and GenCompositor [79]. We adopt the VBench [30] evaluation protocol, reporting Subject Consistency, Background Consistency, Motion Smooth, and FVD to comprehensively assess the quality of the composited videos.

Fig. 6: Harmonization Results. Our method synthesizes shadows and reflections alongside color adaptation.

Table 3: Quantitative comparisons. *Left:* V2V task using VBench metrics against recent video compositing and editing approaches. *Right:* Video Harmonization comparison of lighting and shadow adaptation quality.

Method	Sub. Cons.↑	BG. Cons.↑	Motion↑	FVD↓
AnyV2V [38]	88.02%	92.90%	96.85%	983.56
VACE [31]	89.51%	92.63%	98.21%	942.52
GenCompositor [79]	89.75%	93.43%	98.69%	535.71
Ours	**91.20%**	**93.85%**	**98.71%**	**468.30**

Method	PSNR↑	SSIM↑	LPIPS↓	CLIP↑
Harmonizer [34]	39.8	0.942	0.041	0.961
VTT [22]	40.1	0.931	0.046	0.956
GenCompositor [79]	<u>42.0</u>	<u>0.949</u>	<u>0.039</u>	<u>0.971</u>
Ours	**42.5**	**0.952**	**0.037**	**0.974**

Results and Analysis. Table 3 displays the results. Video editing methods, which often prioritize texture editing, may sometimes distort motion patterns or fail to blend the subject seamlessly. FlexComposer achieves the highest scores in Subject Consistency and Background Consistency, outperforming GenCompositor. As shown in Figure 5, we demonstrate compositing results on two scenarios. Baseline methods exhibit various artifacts: AnyV2V occasionally loses subject coherence or introduces background inconsistencies, VACE struggles with temporal stability across frames, and GenCompositor sometimes alters the object's natural gait pattern or produces misaligned trajectories (see yellow bounding boxes in right columns). In contrast, our method maintains both the intrinsic motion dynamics and precise trajectory adherence throughout the sequence, while seamlessly harmonizing the subject with the target background.

4.4 Video Harmonization

This task evaluates the adaptability of foreground lighting and shadows to the new background environment.

Setup. We use the HYouTube [46] dataset, which contains paired unharmonized and harmonized examples. We compare against Harmonizer [34], VTT [22] and Gencompositor [79], which are specialized for harmonization. We report PSNR, SSIM [71], LPIPS, and CLIP Score [24] to evaluate the fidelity.

Results and Analysis. As shown in Table 3, FlexComposer demonstrates competitive performance across different metrics. Figure 6 presents two representative lighting scenarios: a coastal scene with warm sunset illumination (top) and

Table 4: Ablation Study. Quantitative analysis of key components. We evaluate across different ablation variants.

Variant	EPE ↓	FVD ↓	Motion ↑	LPIPS ↓	CLIP ↑
w/o Canonical Repr.	4.87	145.3	95.1	0.068	0.945
w/o Static Expansion	2.94	128.7	96.8	0.042	0.968
w/o Noise in Static Exp.	2.56	118.2	96.3	0.039	0.969
w/o Relighting Aug.	2.61	106.5	98.1	0.058	0.952
w/o Visibility Gate	2.83	92.4	97.6	0.041	0.969
w/o Latent Transport	5.68	138.9	94.5	0.072	0.938
Curriculum Variants					
Synthetic Only	2.14	154.2	98.4	0.061	0.948
Real Only	3.82	102.6	97.3	0.045	0.965
w/o Generative Data	2.53	98.1	98.2	0.043	0.966
Full Model	**2.39**	**79.8**	**98.8**	**0.036**	**0.975**

Fig. 7: Visual Ablation Analysis. Impact of noise injection in static image expansion.

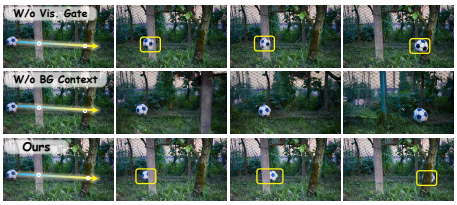

Fig. 8: Visual Ablation on Context and Visibility. Impact of the Background Context and the visible gate in latent injection module.

a beach environment with directional lighting (bottom). Specialized harmonization methods effectively adjust color tones to match ambient lighting conditions, showing their strength in appearance adaptation. Our method attempts to complement color harmonization with geometric cues such as cast shadows. While each method has its merits in different aspects of harmonization, our approach aims to balance both photometric and geometric consistency for composite video generation.

4.5 Ablation Study

To validate the effectiveness of our design choices, we conduct a comprehensive ablation study. We remove key components and analyze their impact on trajectory accuracy (EPE), temporal consistency (FVD), motion preservation (Motion Score from VBench), and harmonization quality (LPIPS, CLIP Score). Results are summarized in Table 4. More details can be found in supplementary materials.

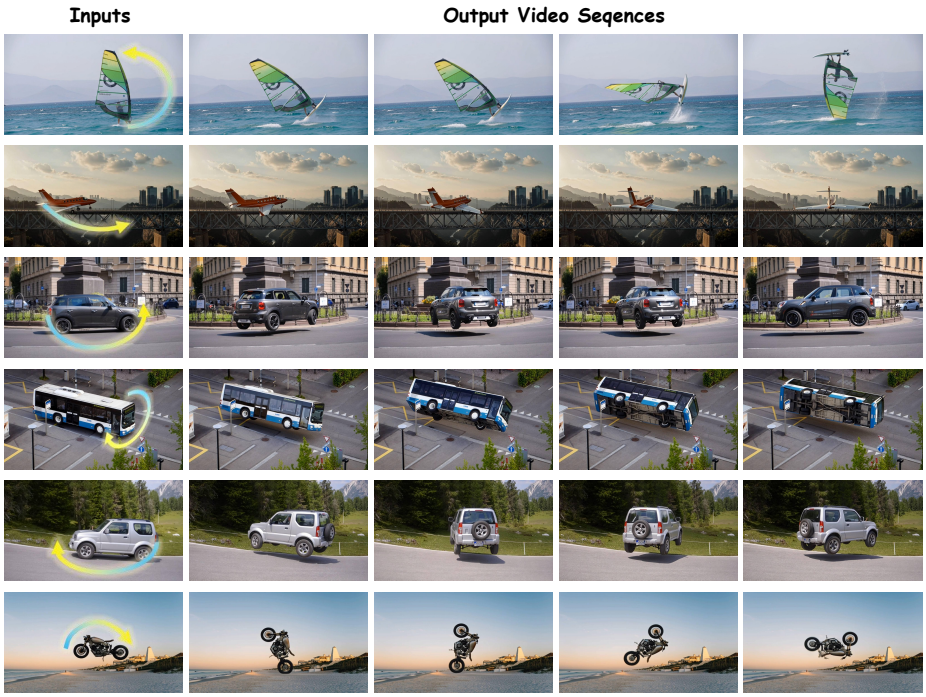

Fig. 9: Qualitative results of complex 3D rotations. FlexComposer synthesizes challenging out-of-plane motions (e.g., flips, rolls) by applying explicit 3D transformations to a SAM3D-reconstructed proxy. Injecting the resulting trajectory ensures temporally consistent synthesis even during extreme pose changes.

Canonical Representation & Noise. Removing the canonical pipeline (*w/o Canonical Repr.*) causes severe motion conflicts, degrading trajectory adherence. For static inputs, disabling noise injection (*w/o Static Noise*) results in rigid, statue-like outputs. As shown in Figure 7, noise is essential for synthesizing plausible micro-motions.

Latent Injection & Visibility. This core module governs the precise placement and integration of the asset into the scene, relying on three components. First, the Latent Transport mechanism aligns canonical features with the trajectory. Without it (*w/o Latent Transport*), sliding artifacts emerge as objects fail to anchor to the scene geometry. Second, as shown in Figure 8 (middle), omitting this context (*w/o BG Context*) compromises environmental stability, leading to severe hallucinations in non-edited regions. Finally, the Visibility Gate manages the interactions; removing it (*w/o Vis. Gate*) results in ghosting artifacts (Figure 8, top), where objects incorrectly blend with occluders rather than disappearing behind them.

Relighting Augmentation & Curriculum. Removing the relighting module leads to a sticker effect with flat lighting. While trajectory metrics remain strong,

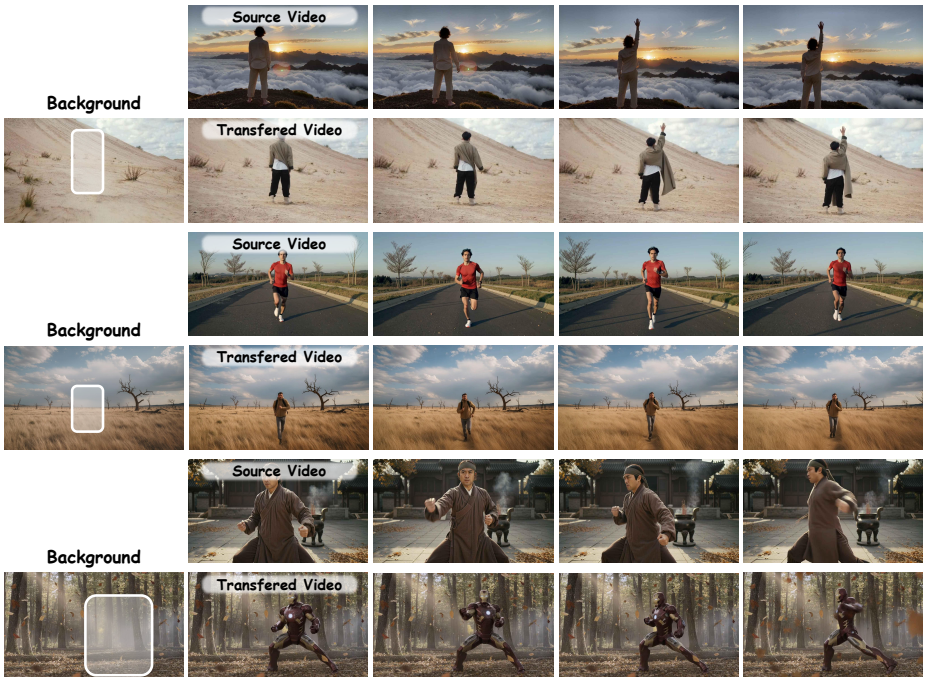

Fig. 10: Qualitative results of motion transfer. By mapping source trajectories to target characters via latent transport, our framework facilitates motion synthesis in new background. This approach supports non-pixel-aligned transfer, allowing for motion transfer across different character and backgrounds.

the perceptual quality significantly degrades as objects appear inconsistent with the target scene lighting. Regarding data, *Synthetic Only* yields precise control but poor realism, while *Real Only* suffers from noisy trajectory signals. Our hybrid curriculum effectively balances geometric precision with photorealistic harmonization.

4.6 Downstream Applications

FlexComposer's unified representation and flexible trajectory control naturally enable advanced video editing tasks beyond standard compositing.

Complex 3D Motion Synthesis. Our framework can synthesize challenging out-of-plane motions, such as flips, rolls, and axial rotations (see Figure 9). We first reconstruct a 3D proxy of the target asset using SAM3D. By applying explicit 3D transformations to this proxy, we extract a precise 2D motion guidance trajectory. Injecting this signal ensures temporally consistent synthesis even during extreme pose changes.

Trajectory-Driven Motion Transfer. By decoupling intrinsic asset features from global spatial coordinates, FlexComposer intrinsically supports non-pixel-

aligned motion transfer (see Figure 10). We can extract the trajectory from source video and seamlessly map it to an entirely different target subject in a new background. This achieves robust motion retargeting across diverse subjects and scenes without requiring rigid alignment or specific fine-tuning.

5 Conclusions & Limitations

We presented **FlexComposer**, a unified framework that bridges the gap between precise geometric controllability and photorealistic video compositing. By decoupling motion from appearance via our Canonical Representation and Latent Transport, FlexComposer achieves precise trajectory adherence for both static and dynamic assets while maintaining high visual fidelity. Extensive experiments demonstrate that our method effectively balances motion precision with photometric harmonization, establishing a robust foundation for controllable video synthesis. Despite these advancements, practical challenges remain to be addressed. While lighting is harmonized, physical interactions are not simulated and rely solely on diffusion priors, leading to implausible physics. Besides, generating long-duration videos with extreme viewpoint changes can suffer from inconsistency. Future work will explore integrating lightweight physics engines to guide the diffusion process and leveraging 3D representations to enhance consistency and temporal stability.

References

1. Adobe: After effects. https://www.adobe.com/products/aftereffects.html (2025), accessed: 2025-01-03 4
2. Adobe Systems Inc.: Mixamo. https://www.mixamo.com (2025) 7
3. Bahmani, S., Shen, T., Ren, J., Huang, J., Jiang, Y., Turki, H., Tagliasacchi, A., Lindell, D.B., Gojcic, Z., Fidler, S., et al.: Lyra: Generative 3d scene reconstruction via video diffusion model self-distillation. arXiv preprint arXiv:2509.19296 (2025) 4
4. Bai, J., Xia, M., Fu, X., Wang, X., Mu, L., Cao, J., Liu, Z., Hu, H., Bai, X., Wan, P., et al.: Recammaster: Camera-controlled generative rendering from a single video. arXiv preprint arXiv:2503.11647 (2025) 4
5. Bai, J., Xia, M., Wang, X., Yuan, Z., Fu, X., Liu, Z., Hu, H., Wan, P., Zhang, D.: Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints. arXiv preprint arXiv:2412.07760 (2024) 4
6. Bian, Y., Xue, Z., Zhang, S., Zhang, S., Jin, W., Li, Y., Zhuang, J., Li, H., Huang, J., Huang, H., et al.: Echo-infinity: Learning evolving memory for real-time infinite video generation. arXiv preprint arXiv:2606.04527 (2026) 3
7. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023) 3
8. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023) 4

9. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., et al.: Video generation models as world simulators. OpenAI Technical Report (2024), https://openai.com/research/video-generation-models-as-world-simulators 3
10. Cai, Y., Zhang, H., Chen, X., Xing, J., Hu, Y., Zhou, Y., Zhang, K., Zhang, Z., Kim, S.Y., Wang, T., et al.: Omnivcus: Feedforward subject-driven video customization with multimodal control conditions. arXiv preprint arXiv:2506.23361 (2025) 2, 4
11. Chen, J., He, T., Fu, Z., Wan, P., Gai, K., Ye, W.: Vino: A unified visual generator with interleaved omnimodal context. arXiv preprint arXiv:2601.02358 (2026) 2, 4
12. Cheng, J., Xiao, T., He, T.: Consistent video-to-video transfer using synthetic dataset. arXiv preprint arXiv:2311.00213 (2023) 4
13. Chu, R., He, Y., Chen, Z., Zhang, S., Xu, X., Xia, B., Wang, D., Yi, H., Liu, X., Zhao, H., et al.: Wan-move: Motion-controllable video generation via latent trajectory guidance. arXiv preprint arXiv:2512.08765 (2025) 4, 8, 9, 10
14. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021) 3
15. Foundry: Nuke. https://www.foundry.com/products/nuke (2025), accessed: 2025-01-03 4
16. Gao, C., Saraf, A., Kopf, J., Huang, J.B.: Dynamic view synthesis from dynamic monocular video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5712–5721 (2021) 4
17. Geng, D., Herrmann, C., Hur, J., Cole, F., Zhang, S., Pfaff, T., Lopez-Guevara, T., Doersch, C., Aytar, Y., Rubinstein, M., et al.: Motion prompting: Controlling video generation with motion trajectories. arXiv preprint arXiv:2412.02700 (2024) 4, 6
18. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014) 3
19. Greff, K., Belletti, F., Beyer, L., Doersch, C., Du, Y., Duckworth, D., Fleet, D.J., Gnanapragasam, D., Golemo, F., Herrmann, C., et al.: Kubric: A scalable dataset generator. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3749–3761 (2022) 7
20. Gu, Z., Yan, R., Lu, J., Li, P., Dou, Z., Si, C., Dong, Z., Liu, Q., Lin, C., Liu, Z., et al.: Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–12 (2025) 4, 6
21. Guo, Y., Yang, C., Rao, A., Agrawala, M., Lin, D., Dai, B.: Sparsectrl: Adding sparse controls to text-to-video diffusion models. In: ECCV. pp. 330–348 (2024) 4
22. Guo, Z., Han, X., Zhang, J., Shan, S., Zheng, H.: Video harmonization with triplet spatio-temporal variation patterns. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19177–19186 (2024) 2, 4, 11
23. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for text-to-video generation. arXiv preprint arXiv:2404.02101 (2024) 4
24. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning (2022), https://arxiv.org/abs/2104.08718 11
25. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017) 9

26. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS (2020) 3
27. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in neural information processing systems **35**, 8633–8646 (2022) 3
28. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022) 8
29. Huang, H.P., Su, Y.C., Sun, D., Jiang, L., Jia, X., Zhu, Y., Yang, M.H.: Fine-grained controllable video generation via object appearance and context. arXiv preprint arXiv:2312.02919 (2023) 4
30. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024) 10
31. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17191–17202 (2025) 2, 4, 10, 11
32. Ju, X., Wang, T., Zhou, Y., Zhang, H., Liu, Q., Zhao, N., Zhang, Z., Li, Y., Cai, Y., Liu, S., et al.: Editverse: Unifying image and video editing and generation with in-context learning. arXiv preprint arXiv:2509.20360 (2025) 2, 4
33. Karsch, K., Sunkavalli, K., Hadap, S., Carr, N., Jin, H., Fonte, R., Sittig, M., Forsyth, D.: Automatic scene inference for 3d object compositing. ACM Transactions on Graphics (TOG) **33**(3), 1–15 (2014) 2, 4
34. Ke, Z., Sun, C., Zhu, L., Xu, K., Lau, R.W.: Harmonizer: Learning to perform white-box image and video harmonization. In: European conference on computer vision. pp. 690–706. Springer (2022) 2, 11
35. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023) 5
36. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024) 3
37. Kong, X., Zhang, Z., Guo, Y., Zhao, Z., Zhang, S., Rao, A.: Composing concepts from images and videos via concept-prompt binding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14800–14810 (2026) 2, 4
38. Ku, M., Wei, C., Ren, W., Yang, H., Chen, W.: Anyv2v: A tuning-free framework for any video-to-video editing tasks. arXiv preprint arXiv:2403.14468 (2024) 4, 10, 11
39. Kuaishou: Keling (2024), https://kling.kuaishou.com/ 3
40. Lee, Y.C., Zhang, Z., Blackburn-Matzen, K., Niklaus, S., Zhang, J., Huang, J.B., Liu, F.: Fast view synthesis of casual videos with soup-of-planes. In: European Conference on Computer Vision. pp. 278–296. Springer (2024) 4
41. Lei, G., Wang, C., Li, H., Zhang, R., Wang, Y., Xu, W.: Animateanything: Consistent and controllable animation for video generation. arXiv preprint arXiv:2411.10836 (2024) 4
42. Li, Q., Xing, Z., Wang, R., Zhang, H., Dai, Q., Wu, Z.: Magicmotion: Controllable video generation with dense-to-sparse trajectory guidance. arXiv preprint arXiv:2503.16421 (2025) 9, 10

43. Li, Y., Wang, X., Zhang, Z., Wang, Z., Yuan, Z., Xie, L., Zou, Y., Shan, Y.: Image conductor: Precision control for interactive video synthesis. arXiv preprint arXiv:2406.15339 (2024) 9, 10
44. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022) 7
45. Liu, R., Yuan, H., Dong, B., Xing, J., Wang, J., Zhao, R., Xing, Y., Chen, W., Wang, F.: Unilumos: Fast and unified image and video relighting with physics-plausible feedback. arXiv preprint arXiv:2511.01678 (2025) 6
46. Lu, X., Huang, S., Niu, L., Cong, W., Zhang, L.: Deep video harmonization with color mapping consistency. arXiv preprint arXiv:2205.00687 (2022) 11
47. Ma, W.D.K., Lewis, J.P., Kleijn, W.B.: Trailblazer: Trajectory control for diffusion-based video generation. In: SIGGRAPH Asia (2024) 4
48. Ma, Y., Feng, K., Zhang, X., Liu, H., Zhang, D.J., Xing, J., Zhang, Y., Yang, A., Wang, Z., Chen, Q.: Follow-your-creation: Empowering 4d creation through video inpainting. arXiv preprint arXiv:2506.04590 (2025) 4
49. Ma, Y., He, Y., Cun, X., Wang, X., Chen, S., Li, X., Chen, Q.: Follow your pose: Pose-guided text-to-video generation using pose-free videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4117–4125 (2024) 4
50. Mai, J., Wang, C., Qian, G.G., Menapace, W., Tulyakov, S., Ghanem, B., Wonka, P., Mirzaei, A.: Easyv2v: A high-quality instruction-based video editing framework. arXiv preprint arXiv:2512.16920 (2025) 4
51. Mou, C., Cao, M., Wang, X., Zhang, Z., Shan, Y., Zhang, J.: Revideo: Remake a video with motion and content control. arXiv preprint arXiv:2405.13865 (2024) 4
52. Namekata, K., Bahmani, S., Wu, Z., Kant, Y., Gilitschenski, I., Lindell, D.B.: Sg-i2v: Self-guided trajectory control in image-to-video generation. arXiv preprint arXiv:2411.04989 (2024) 4
53. Pandey, K., Gadelha, M., Hold-Geoffroy, Y., Singh, K., Mitra, N.J., Guerrero, P.: Motion modes: What could happen next? arXiv preprint arXiv:2412.00148 (2024) 4
54. Peebles, W., Xie, S.: Scalable diffusion models with transformers (2023), https://arxiv.org/abs/2212.09748 3
55. Porter, T., Duff, T.: Compositing digital images. ACM SIGGRAPH Computer Graphics **18**(3), 253–259 (1984) 4
56. Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: Fatezero: Fusing attentions for zero-shot text-based video editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15932–15942 (2023) 4
57. Ren, X., Shen, T., Huang, J., Ling, H., Lu, Y., Nimier-David, M., Müller, T., Keller, A., Fidler, S., Gao, J.: Gen3c: 3d-informed world-consistent video generation with precise camera control. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6121–6132 (2025) 4
58. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022) 3
59. Sun, W., Chen, S., Liu, F., Chen, Z., Duan, Y., Zhu, J., Zhang, J., Wang, Y.: Dimensionx: Create any 3d and 4d scenes from a single image with decoupled video diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13695–13706 (2025) 4
60. Team, K., Chen, J., Ci, Y., Du, X., Feng, Z., Gai, K., Guo, S., Han, F., He, J., He, K., et al.: Kling-omni technical report. arXiv preprint arXiv:2512.16776 (2025) 2, 4

61. Tu, Y., Luo, H., Chen, X., Ji, S., Bai, X., Zhao, H.: Videoanydoor: High-fidelity video object insertion with precise motion control. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–11 (2025) 2, 4
62. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018) 9
63. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025) 6, 8
64. Wang, C., Zhuang, P., Ngo, T.D., Menapace, W., Siarohin, A., Vasilkovsky, M., Skorokhodov, I., Tulyakov, S., Wonka, P., Lee, H.Y.: 4real-video: Learning generalizable photo-realistic 4d video diffusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17723–17732 (2025) 4
65. Wang, H., Ouyang, H., Wang, Q., Wang, W., Cheng, K.L., Chen, Q., Shen, Y., Wang, L.: Levitor: 3d trajectory oriented image-to-video synthesis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12490–12500 (2025) 9, 10
66. Wang, H., Ouyang, H., Wang, Q., Yu, Y., Meng, Y., Wang, W., Cheng, K.L., Ma, S., Bai, Q., Li, Y., et al.: The world is your canvas: Painting promptable events with reference images, trajectories, and text. arXiv preprint arXiv:2512.16924 (2025) 4
67. Wang, J., Zhang, Y., Zou, J., Zeng, Y., Wei, G., Yuan, L., Li, H.: Boximator: Generating rich and controllable motions for video synthesis. arXiv preprint arXiv:2402.01566 (2024) 4
68. Wang, Q., Ye, V., Gao, H., Zeng, W., Austin, J., Li, Z., Kanazawa, A.: Shape of motion: 4d reconstruction from a single video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9660–9672 (2025) 4
69. Wang, W., Song, H., Zhao, S., Shen, J., Zhao, S., Hoi, S.C., Ling, H.: Learning unsupervised video object segmentation through visual attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3064–3074 (2019) 9
70. Wang, X., Yuan, H., Zhang, S., Chen, D., Wang, J., Zhang, Y., Shen, Y., Zhao, D., Zhou, J.: Videocomposer: Compositional video synthesis with motion controllability. Advances in Neural Information Processing Systems **36** (2024) 4
71. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004) 9, 11
72. Wang, Z., Yuan, Z., Wang, X., Li, Y., Chen, T., Xia, M., Luo, P., Shan, Y.: Motionctrl: A unified and flexible motion controller for video generation. In: SIGGRAPH (2024) 4
73. Wei, C., Liu, Q., Ye, Z., Wang, Q., Wang, X., Wan, P., Gai, K., Chen, W.: Univideo: Unified understanding, generation, and editing for videos. arXiv preprint arXiv:2510.08377 (2025) 2
74. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7623–7633 (2023) 4
75. Wu, R., Gao, R., Poole, B., Trevithick, A., Zheng, C., Barron, J.T., Holynski, A.: Cat4d: Create anything in 4d with multi-view video diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26057–26068 (2025) 4

76. Wu, Y., Chen, L., Li, R., Wang, S., Xie, C., Zhang, L.: Insvie-1m: Effective instruction-based video editing with elaborate dataset construction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16692–16701 (2025) 4
77. Xian, W., Huang, J.B., Kopf, J., Kim, C.: Space-time neural irradiance fields for free-viewpoint video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9421–9431 (2021) 4
78. Xiao, Y., Wang, J., Xue, N., Karaev, N., Makarov, Y., Kang, B., Zhu, X., Bao, H., Shen, Y., Zhou, X.: Spatialtrackerv2: 3d point tracking made easy. ArXiv **abs/2507.12462** (2025) 5
79. Yang, S., Li, X., Cun, X., Wang, G., Li, L., Shan, Y., Zhang, J.: Gencompositor: generative video compositing with diffusion transformer. arXiv preprint arXiv:2509.02460 (2025) 4, 10, 11
80. Yang, X., Xie, J., Yang, Y., Huang, Y., Xu, M., Wu, Q.: Unified video editing with temporal reasoner. arXiv preprint arXiv:2512.07469 (2025) 2, 4
81. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024) 3
82. Ye, Z., He, X., Liu, Q., Wang, Q., Wang, X., Wan, P., Zhang, D., Gai, K., Chen, Q., Luo, W.: Unic: Unified in-context video editing. arXiv preprint arXiv:2506.04216 (2025) 2, 4
83. Yin, S., Wu, C., Liang, J., Shi, J., Li, H., Ming, G., Duan, N.: Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. arXiv preprint arXiv:2308.08089 (2023) 4
84. You, M., Zhu, Z., Liu, H., Hou, J.: Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. arXiv preprint arXiv:2405.15364 (2024) 4
85. YU, M., Hu, W., Xing, J., Shan, Y.: Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. arXiv preprint arXiv:2503.05638 (2025) 4
86. Yu, S., Fang, J.Z., Zheng, J., Sigurdsson, G., Ordonez, V., Piramuthu, R., Bansal, M.: Zero-shot controllable image-to-video animation via motion decomposition. In: ACM MM (2024) 4
87. Yu, W., Xing, J., Yuan, L., Hu, W., Li, X., Huang, Z., Gao, X., Wong, T.T., Shan, Y., Tian, Y.: Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. arXiv preprint arXiv:2409.02048 (2024) 4
88. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023) 3
89. Zhang, S., Xu, H., Guo, S., Xie, Z., Bao, H., Xu, W., Zou, C.: Spatialcrafter: Unleashing the imagination of video diffusion models for scene reconstruction from limited observations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27794–27805 (2025) 2, 4
90. Zhang, S., Xue, Z., Fu, S., Huang, J., Kong, X., Ma, Y., Huang, H., Duan, N., Rao, A.: Astrolabe: Steering forward-process reinforcement learning for distilled autoregressive video models. arXiv preprint arXiv:2603.17051 (2026) 3
91. Zhang, S., Zhang, Y., Zheng, Q., Ma, R., Hua, W., Bao, H., Xu, W., Zou, C.: 3d-scenedreamer: Text-driven 3d-consistent scene generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10170–10180 (2024) 2, 4

92. Zhang, S., Zhao, C.: Pragmatist: Multiview conditional diffusion models for high-fidelity 3d reconstruction from unposed sparse views. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 10112–10120 (2025) 2, 4
93. Zhang, Z., Liao, J., Li, M., Dai, Z., Qiu, B., Zhu, S., Qin, L., Wang, W.: Tora: Trajectory-oriented diffusion transformer for video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2063–2073 (2025) 4, 9, 10
94. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all (March 2024), https://github.com/hpcaitech/Open-Sora 3
95. Zhou, J., Gao, H., Voleti, V., Vasishta, A., Yao, C.H., Boss, M., Torr, P., Rupprecht, C., Jampani, V.: Stable virtual camera: Generative view synthesis with diffusion models. arXiv preprint arXiv:2503.14489 (2025) 4