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

ANONYMOUS AUTHOR(S)
SUBMISSION ID: 351

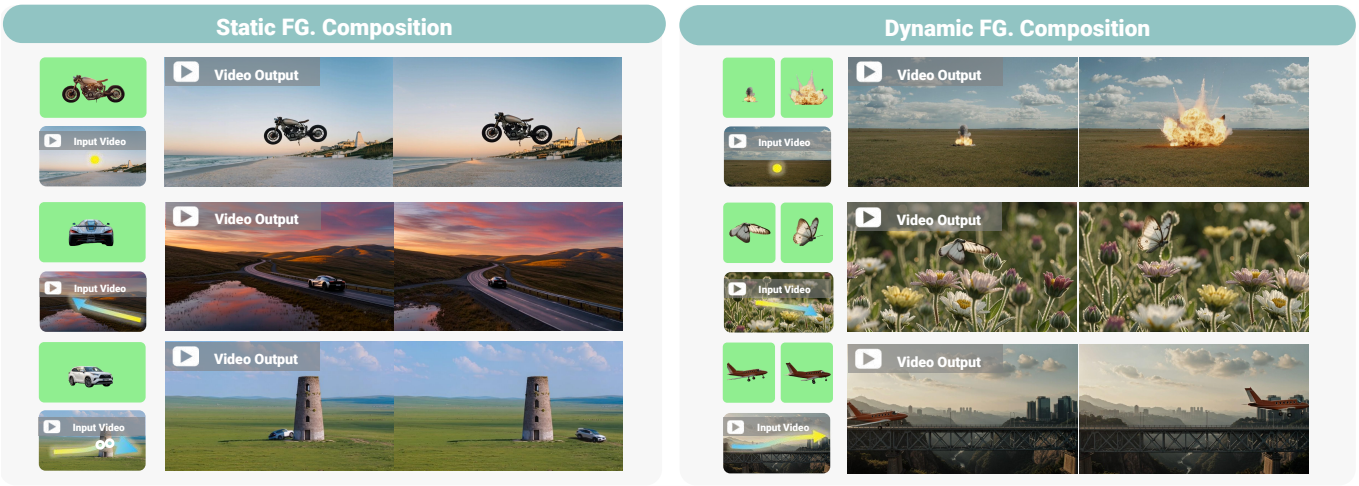

Fig. 1. **Versatile Video Compositing with FlexComposer.** Our framework unifies the insertion of diverse assets into target videos through user-defined trajectories (visualized as arrows in input thumbnails). **Left (Static Foreground Comp.):** Given a single static image, the model synthesizes realistic 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.

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 robustly 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 significantly outperforms state-of-the-art methods in visual quality, temporal consistency, and trajectory adherence.

CCS Concepts: • **Computing methodologies** → **Imaging/Video**.

Additional Key Words and Phrases: Video compositing; Video Diffusion Model; Trajectory control

ACM Reference Format:
Anonymous Author(s). 2026. FlexComposer: Unified Video Compositing from Images to Dynamic Footage with Flexible Trajectory Control. *ACM Trans. Graph.* 1, 1 (January 2026), 16 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM 0730-0301/2026/1-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

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, 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 [Karsch et al. 2014] and harmonization techniques [Guo et al. 2024a; Ke et al. 2022] 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 [Cai et al. 2025; Chen et al. 2026; Jiang et al. 2025; Ju et al. 2025; Team et al. 2025; Tu et al. 2025; Yang et al. 2025b; Ye et al. 2025] leverage diffusion models to address these limitations. These approaches offer promising avenues for composition, ranging from zero-shot subject animators [Cai et al. 2025; Tu et al. 2025] to unified in-context editors [Chen et al. 2026; Ju et al. 2025; Wei et al. 2025; Ye et al. 2025] and large-scale foundation models [Team et al. 2025; Yang et al. 2025b]. However, despite their versatility, a critical gap remains: most methods either hallucinate motion from static references [Tu et al. 2025] or operate at abstract semantic levels [Ju et al. 2025; Wei et al. 2025]. 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 (Figure 2, bottom). 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 [Zhang et al. 2023]), 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 [Goodfellow et al. 2014] and autoregressive transformers [Esser et al. 2021] to Diffusion Models (VDMs) [Blattmann et al. 2023; Ho et al. 2020, 2022], which now serve as the gold standard for high-fidelity generation. While Latent Diffusion Models (LDMs) [Rombach et al. 2022] balanced quality and efficiency, the recent emergence of Video Diffusion Transformers (DiTs) [Kong et al. 2024; Peebles and Xie 2023; Yang et al. 2024], exemplified by Sora [Brooks et al. 2024; Kuaishou 2024; Zheng et al. 2024], 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.

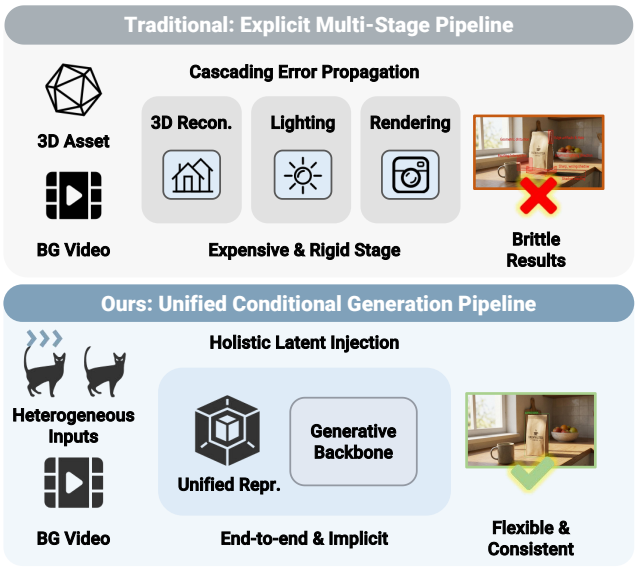

Fig. 2. **Motivation.** Traditional explicit pipelines (top) suffer from error accumulation across decoupled stages. In contrast, our **FlexComposer** (bottom) employs a unified conditional generation framework, implicitly handling geometry and illumination to achieve robust, flexible harmonization end-to-end.

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 [Guo et al. 2024b; Huang et al. 2023; Mou et al. 2024; Namekata et al. 2024; Pandey et al. 2024; Wang et al. 2024a; Yin et al. 2023; Yu et al. 2024a], with recent works integrating trajectory adapters directly into DiT architectures [Chu et al. 2025; Geng et al. 2024; Ma et al. 2024b; Wang et al. 2025b, 2024c; Zhang et al. 2025]. 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. Approaches range from traditional reconstruction-based methods [Gao et al. 2021; Lee et al. 2024; Wang et al. 2025c; Xian et al. 2021] to generative techniques using multi-view diffusion, inpainting, or pose conditioning [Bahmani et al. 2025; Bai et al. 2025, 2024; He et al. 2024; Ma et al. 2025; Ren et al. 2025; Sun et al. 2025; Wang et al. 2025d; Wu et al. 2025b;

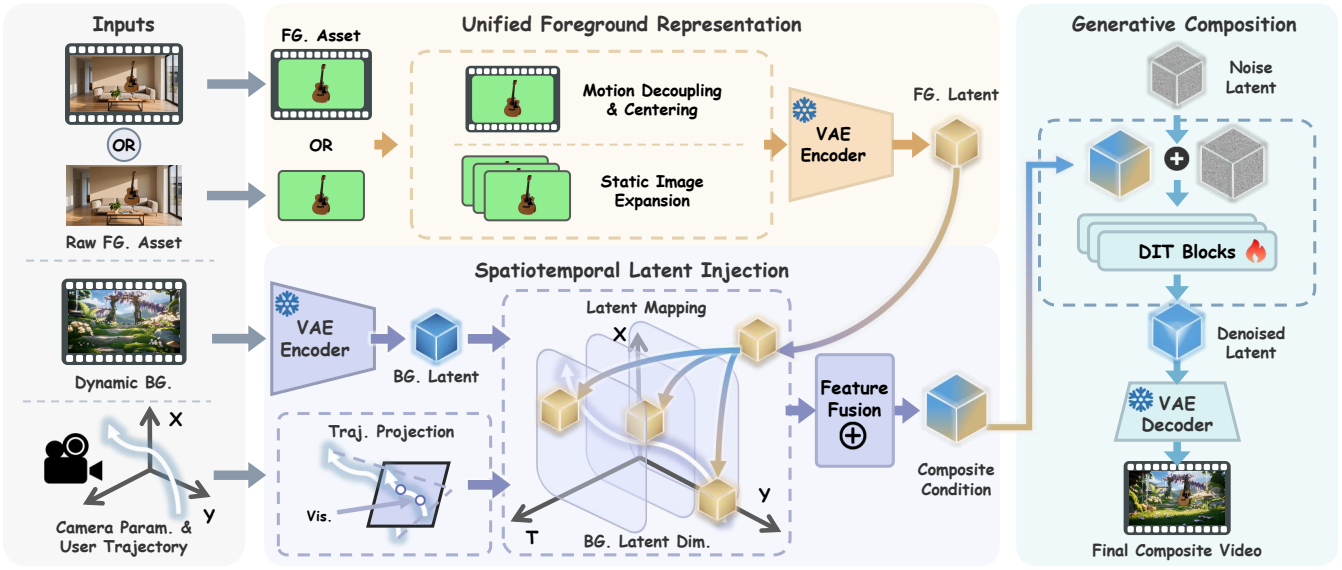

Fig. 3. The Pipeline of FlexComposer. Our framework enables controllable insertion of diverse assets into dynamic videos. (1) **Inputs**: The system accepts a dynamic background video, and either a static image or dynamic video as the foreground asset, along with user-defined 3D trajectory and visibility masks. (2) **Unified Representation**: Dynamic assets are stabilized, and static assets are temporally expanded with noise. Both are encoded into a unified canonical feature space, preserving identity and local motion. (3) **Spatiotemporal Latent Injection**: The core module warps the canonical foreground features into the background latent space based on the trajectory. A visibility gate is applied to handle occlusions. (4) **Generative Composition**: The warped features serve as spatial-aware conditions for video diffusion model, which synthesizes the final photorealistic and temporally consistent composite video.

You et al. 2024; YU et al. 2025; Yu et al. 2024b; Zhou et al. 2025]. 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 [Gu et al. 2025; Lei et al. 2024; Ma et al. 2024a; Wang et al. 2024b]. 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 [Karsch et al. 2014] rely on 3D reconstruction, lighting estimation, and ray tracing in separate stages, with error accumulation across stages limiting robustness. Professional tools [Adobe 2025; Foundry 2025] built on Porter and Duff's compositing operator [Porter and Duff 1984] enable non-destructive layer-based workflows but require extensive manual per-frame artistry for masking and effect tuning. Recent lightweight approaches focus on color harmonization [Guo et al. 2024a; Yang et al. 2025a], 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 [Brooks et al. 2023] and has evolved toward video through per-video fine-tuning [Qi et al. 2023; Wu et al. 2023] and inference-only protocols [Cai et al. 2025; Chen et al. 2026; Cheng et al. 2023; Ju et al. 2025; Ku et al. 2024; Mai et al. 2025; Team et al. 2025; Wu et al. 2025a; Yang et al. 2025b; Ye et al. 2025]. While diffusion priors enable semantic editing, they lack precise trajectory mechanisms for asset insertion. Conversely, generative compositing frameworks [Cai et al. 2025; Jiang et al. 2025; Tu et al. 2025] 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 $\mathcal{T}_{user}$. 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. 3, 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 [Kirillov et al. 2023] for

segmentation and SpatialTracker v2 [Xiao et al. 2025] 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 background contexts, we apply random illumination perturbations to the frames via a relighting module [Liu et al. 2025]. 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 [Geng et al. 2024; Gu et al. 2025] 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 [Wan et al. 2025] 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 [Lipman et al. 2022]. Crucially, the velocity prediction network $\mathbf{v}_\theta$ now takes the concatenated latent $\mathbf{Z}_{\text{in}}$ as its spatial input:

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

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 [Adobe Systems Inc. 2025] and synthetic environments [Greff et al. 2022] 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 models. We refine the model for 1k steps on 5k highly diverse clips with pseudo-ground-truth trajectories.

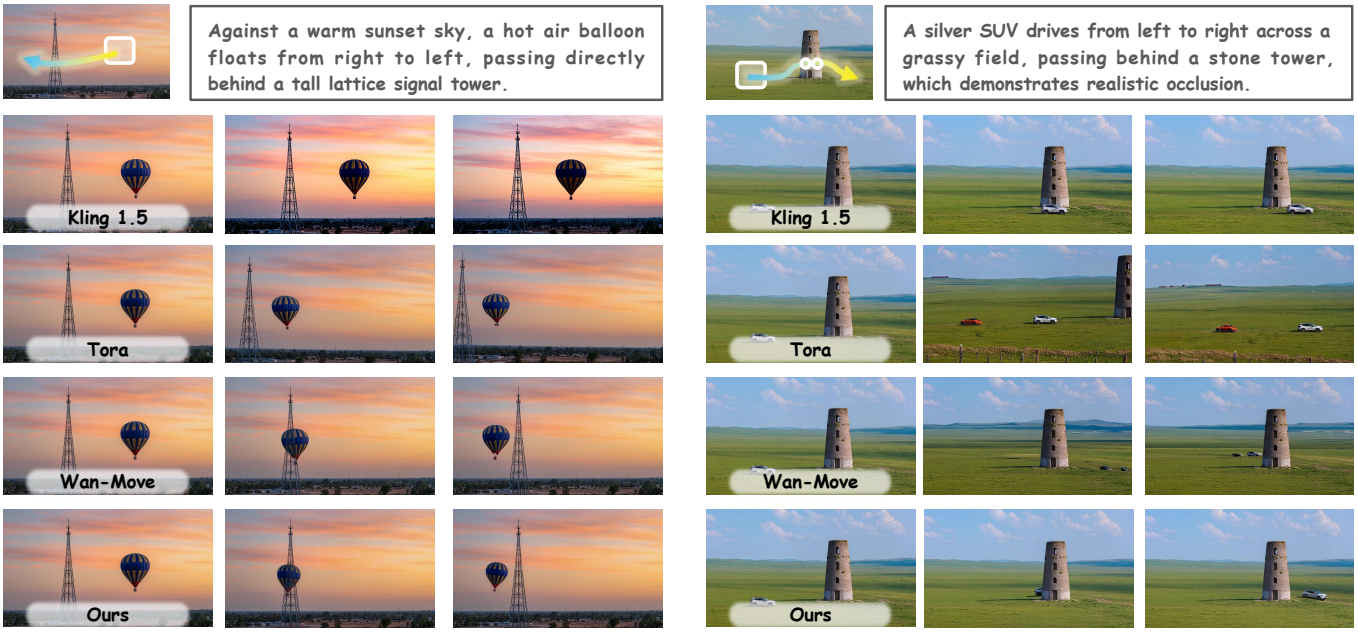

Fig. 4. **Qualitative Comparison on Static Foreground Compositing.** We compare our method against trajectory-controlled I2V methods on two challenging scenarios. Our method demonstrates superior occlusion handling and trajectory adherence while maintaining visual fidelity

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 [Li et al. 2024]	14.92	54.5	512.8	11.55	0.468
LeviTor [Wang et al. 2025a]	3.65	22.3	116.1	13.20	0.515
Tora [Zhang et al. 2025]	3.52	26.1	128.5	13.65	0.492
MagicMotion [Li et al. 2025]	3.48	24.5	112.9	12.90	0.535
Wan-Move [Chu et al. 2025]	2.50	14.7	94.3	16.50	0.610
Ours (I2V)	2.38	14.1	89.5	16.85	0.625
Ours (V2V)	**2.15**	**13.4**	**82.6**	**17.40**	**0.658**

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 [Li et al. 2024]	15.55	34.8	422.5	13.50	0.485
LeviTor [Wang et al. 2025a]	3.45	18.3	99.2	15.55	0.538
MagicMotion [Li et al. 2025]	3.25	17.6	97.1	14.85	0.562
Tora [Zhang et al. 2025]	3.32	22.8	101.5	15.65	0.548
Wan-Move [Chu et al. 2025]	2.60	12.2	83.5	17.80	0.640
Ours (I2V)	2.42	11.9	80.4	17.95	0.655
Ours (V2V)	**2.38**	**11.2**	**74.9**	**18.60**	**0.682**

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 [Wan et al. 2025], initializing the backbone with Wan-Move weights [Chu et al. 2025] 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) [Hu et al. 2022] ($r = 64$) 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.

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 [Wang et al. 2019] and MoveBench [Chu et al. 2025]. Following standard protocols, we evaluate the fidelity of generated videos using FVD [Unterthiner et al. 2018] and FID [Heusel et al. 2017] for perceptual quality, alongside PSNR and SSIM [Wang et al. 2004] for frame-level accuracy. We compare against state-of-the-art motion-controllable methods [Chu et al. 2025; Li et al. 2025, 2024; Wang et al. 2025a; Zhang et al. 2025] under two settings: *I2V* (first frame only) and *V2V* (full background context).

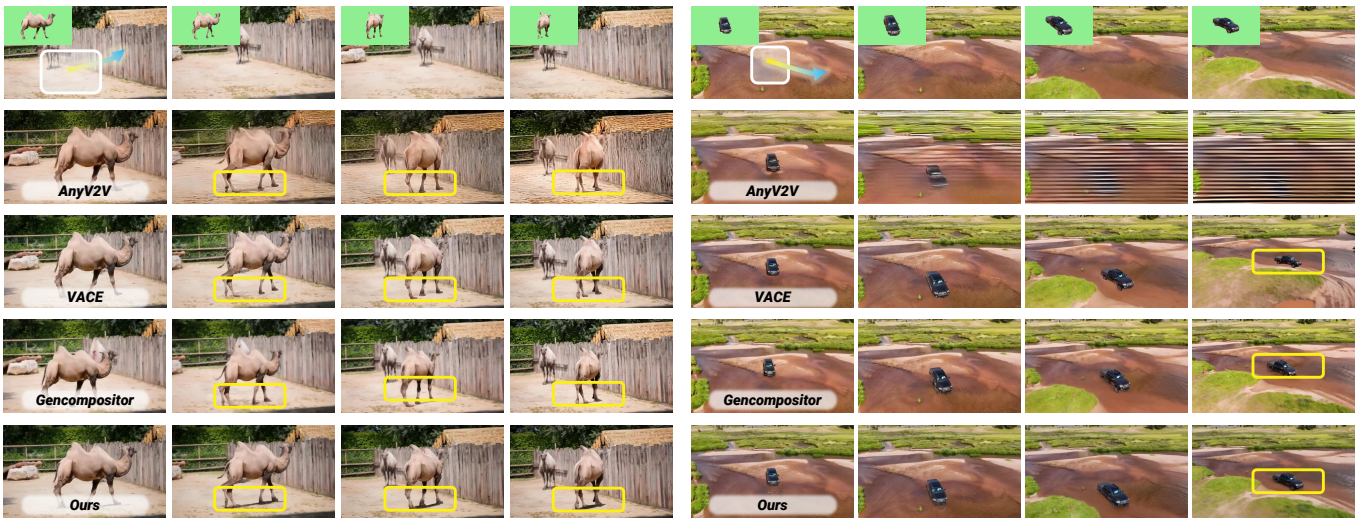

Fig. 5. **Qualitative Comparison on Dynamic Foreground Compositing.** We evaluate our method against video editing and compositing approaches on the V2V task. Our method effectively preserves intrinsic object dynamics while following user-defined trajectories, achieving superior motion fidelity and background consistency compared to baseline methods.

Table 3. **Quantitative comparison on the V2V task using VBench metrics.** Comparision with recent video compositing and editing approaches.

Method	Sub. Cons. ↑	BG. Cons. ↑	Motion ↑	FVD ↓
AnyV2V [Ku et al. 2024]	88.02%	92.90%	96.85%	983.56
VACE [Jiang et al. 2025]	89.51%	92.63%	98.21%	942.52
GenCompositor [Yang et al. 2025a]	89.75%	93.43%	98.69%	535.71
Ours	**91.20%**	**93.85%**	**98.71%**	**468.30**

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

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 [Li et al. 2024] and Tora [Zhang et al. 2025]. These results suggest that our trajectory injection mechanism effectively guides motion generation, offering a robust alternative to pixel-level 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 [Ku et al. 2024], VACE [Jiang et al. 2025], and GenCompositor [Yang et al. 2025a]. We adopt the VBench [Huang et al. 2024] evaluation protocol, reporting Subject Consistency, Background Consistency, Motion Smooth, and FVD to comprehensively assess the quality of the composited videos.

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 [Lu et al. 2022] dataset, which contains paired unharmonized and harmonized examples. We compare against Harmonizer [Ke et al. 2022], VTT [Guo et al. 2024a] and Gencompositor [Yang et al. 2025a], which are specialized for harmonization. We report PSNR, SSIM [Wang et al. 2004], LPIPS, and CLIP Score [Hessel et al. 2022] to evaluate the fidelity.

Table 4. **Video Harmonization.** Comparison of lighting and shadow adaptation quality.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	CLIP ↑
Harmonizer [Ke et al. 2022]	39.8	0.942	0.041	0.961
VTT [Guo et al. 2024a]	40.1	0.931	0.046	0.956
Gencompositor [Yang et al. 2025a]	42.0	0.949	0.039	0.971
Ours	**42.5**	**0.952**	**0.037**	**0.974**

Table 5. **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**

Results and Analysis. As shown in Table 4, FlexComposer demonstrates competitive performance across different metrics. Figure 6 presents two representative lighting scenarios: a coastal scene with warm sunset illumination (top) and 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 5. More details can be found in supplementary materials.

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.

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

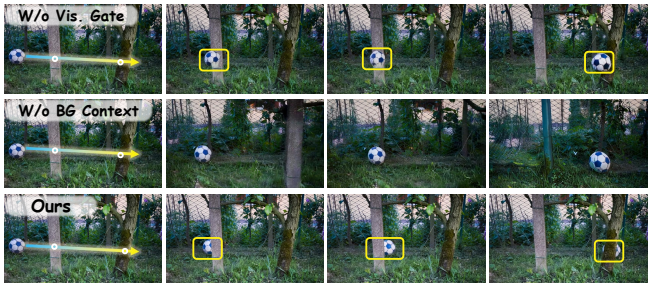

Fig. 8. **Visual Ablation on Context and Visibility.** Impact of the Background Context and the visible gate in latent injection module.

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

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

Adobe. 2025. After Effects. https://www.adobe.com/products/aftereffects.html. Accessed: 2025-01-03.

Adobe Systems Inc. 2025. Mixamo. https://www.mixamo.com.

Sherwin Bahmani, Tianchang Shen, Jiawei Ren, Jiahui Huang, Yifeng Jiang, Haithem Turki, Andrea Tagliasacchi, David B Lindell, Zan Gojcic, Sanja Fidler, et al. 2025. Lyra: Generative 3d scene reconstruction via video diffusion model self-distillation. *arXiv preprint arXiv:2509.19296* (2025).

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. 2025. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647* (2025).

Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Xiao Fu, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. 2024. Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints. *arXiv preprint arXiv:2412.07760* (2024).

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

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18392–18402.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. 2024. Video generation models as world simulators. OpenAI Technical Report. https://openai.com/research/video-generation-models-as-world-simulators

Yuanhao Cai, He Zhang, Xi Chen, Jinbo Xing, Yiwei Hu, Yuqian Zhou, Kai Zhang, Zhifei Zhang, Soo Ye Kim, Tianyu Wang, et al. 2025. OmniVCus: Feedforward Subject-driven Video Customization with Multimodal Control Conditions. *arXiv preprint arXiv:2506.23361* (2025).

Junyi Chen, Tong He, Zhoujie Fu, Pengfei Wan, Kun Gai, and Weicai Ye. 2026. VINO: A Unified Visual Generator with Interleaved OmniModal Context. *arXiv preprint arXiv:2601.02358* (2026).

Jiaxin Cheng, Tianjun Xiao, and Tong He. 2023. Consistent video-to-video transfer using synthetic dataset. *arXiv preprint arXiv:2311.00213* (2023).

Ruihang Chu, Yefei He, Zhekai Chen, Shiwei Zhang, Xiaogang Xu, Bin Xia, Dingdong Wang, Hongwei Yi, Xihui Liu, Hengshuang Zhao, et al. 2025. Wan-move: Motion-controllable video generation via latent trajectory guidance. *arXiv preprint arXiv:2512.08765* (2025).

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12873–12883.

Foundry. 2025. Nuke. https://www.foundry.com/products/nuke. Accessed: 2025-01-03.

Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. 2021. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5712–5721.

Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Carl Doersch, Yusuf Aytar, Michael Rubinstein, et al. 2024. Motion Prompting: Controlling Video Generation with Motion Trajectories. *arXiv preprint arXiv:2412.02700* (2024).

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. 2022. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3749–3761.

Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. 2025. 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*. 1–12.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. 2024b. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *ECCV*. 330–348.

Zonghui Guo, Xinyu Han, Jie Zhang, Shiguang Shan, and Haiyong Zheng. 2024a. Video Harmonization with Triplet Spatio-Temporal Variation Patterns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19177–19186.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. 2024. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101* (2024).

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2022. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. arXiv:2104.08718 [cs.CV] https://arxiv.org/abs/2104.08718

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *NeurIPS* (2020).

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022. Video diffusion models. *Advances in neural information processing systems* 35 (2022), 8633–8646.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.

Hsin-Ping Huang, Yu-Chuan Su, Deqing Sun, Lu Jiang, Xuhui Jia, Yukun Zhu, and Ming-Hsuan Yang. 2023. Fine-grained controllable video generation via object appearance and context. *arXiv preprint arXiv:2312.02919* (2023).

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, 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.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. 2025. VACE: All-in-One Video Creation and Editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 17191–17202.

Xuan Ju, Tianyu Wang, Yuqian Zhou, He Zhang, Qing Liu, Nanxuan Zhao, Zhifei Zhang, Yijun Li, Yuanhao Cai, Shaoteng Liu, et al. 2025. Editverse: Unifying image and video editing and generation with in-context learning. *arXiv preprint arXiv:2509.20360* (2025).

Kevin Karsch, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, Hailin Jin, Rafael Fonte, Michael Sittig, and David Forsyth. 2014. Automatic scene inference for 3d object compositing. *ACM Transactions on Graphics (TOG)* 33, 3 (2014), 1–15.

Zhanghan Ke, Chunyi Sun, Lei Zhu, Ke Xu, and Rynson WH Lau. 2022. Harmonizer: Learning to perform white-box image and video harmonization. In *European conference on computer vision*. Springer, 690–706.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4026.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, et al. 2024. Hunyuanvideo: A Systematic Framework For Large Video Generative Models. *arXiv preprint arXiv:2412.03603* (2024).

Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu Chen. 2024. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468* (2024).

Kuaishou. 2024. Keling. https://kling.kuaishou.com/

Yao-Chih Lee, Erika Lu, Sarah Rumbley, Michal Geyer, Jia-Bin Huang, Tali Dekel, and Forrester Cole. 2025. Generative omnimatte: Learning to decompose video into layers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 12522–12532.

Yao-Chih Lee, Zhoutong Zhang, Kevin Blackburn-Matzen, Simon Niklaus, Jianming Zhang, Jia-Bin Huang, and Feng Liu. 2024. Fast View Synthesis of Casual Videos with Soup-of-Planes. In *European Conference on Computer Vision*. Springer, 278–296.

Guojun Lei, Chi Wang, Hong Li, Rong Zhang, Yikai Wang, and Weiwei Xu. 2024. AnimateAnything: Consistent and Controllable Animation for Video Generation. *arXiv preprint arXiv:2411.10836* (2024).

Quanhao Li, Zhen Xing, Rui Wang, Hui Zhang, Qi Dai, and Zuxuan Wu. 2025. Magicmotion: Controllable video generation with dense-to-sparse trajectory guidance. *arXiv preprint arXiv:2503.16421* (2025).

Yaowei Li, Xintao Wang, Zhaoyang Zhang, Zhouxia Wang, Ziyang Yuan, Liangbin Xie, Yuexian Zou, and Ying Shan. 2024. Image conductor: Precision control for interactive video synthesis. *arXiv preprint arXiv:2406.15339* (2024).

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747* (2022).

Ropeway Liu, Hangjie Yuan, Bo Dong, Jiazheng Xing, Jinwang Wang, Rui Zhao, Yan Xing, Weihua Chen, and Fan Wang. 2025. UniLumos: Fast and Unified Image and Video Relighting with Physics-Plausible Feedback. *arXiv preprint arXiv:2511.01678* (2025).

Xinyuan Lu, Shengyuan Huang, Li Niu, Wenyan Cong, and Liqing Zhang. 2022. Deep video harmonization with color mapping consistency. *arXiv preprint arXiv:2205.00687* (2022).

Wan-Duo Kurt Ma, John P Lewis, and W Bastiaan Kleijn. 2024b. Trailblazer: Trajectory control for diffusion-based video generation. In *SIGGRAPH Asia*.

Yue Ma, Kunyu Feng, Xinhua Zhang, Hongyu Liu, David Junhao Zhang, Jinbo Xing, Yinhan Zhang, Ayden Yang, Zeyu Wang, and Qifeng Chen. 2025. Follow-Your-Creation: Empowering 4D Creation through Video Inpainting. *arXiv preprint arXiv:2506.04590* (2025).

Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. 2024a. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 4117–4125.

Jinjie Mai, Chaoyang Wang, Guocheng Gordon Qian, Willi Menapace, Sergey Tulyakov, Bernard Ghanem, Peter Wonka, and Ashkan Mirzaei. 2025. EasyV2V: A High-quality

Instruction-based Video Editing Framework. *arXiv preprint arXiv:2512.16920* (2025).
Chong Mou, Mingdeng Cao, Xintao Wang, Zhaoyang Zhang, Ying Shan, and Jian Zhang. 2024. ReVideo: Remake a Video with Motion and Content Control. *arXiv preprint arXiv:2405.13865* (2024).
Koichi Namekata, Sherwin Bahmani, Ziyi Wu, Yash Kant, Igor Gilitschenski, and David B Lindell. 2024. Sg-i2v: Self-guided trajectory control in image-to-video generation. *arXiv preprint arXiv:2411.04989* (2024).
Karran Pandey, Matheus Gadelha, Yannick Hold-Geoffroy, Karan Singh, Niloy J Mitra, and Paul Guerrero. 2024. Motion Modes: What Could Happen Next? *arXiv preprint arXiv:2412.00148* (2024).
William Peebles and Saining Xie. 2023. Scalable Diffusion Models with Transformers. arXiv:2212.09748 [cs.CV] https://arxiv.org/abs/2212.09748
Thomas Porter and Tom Duff. 1984. Compositing digital images. *ACM SIGGRAPH Computer Graphics* 18, 3 (1984), 253–259.
Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. 2023. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15932–15942.
Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. 2024. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. *ArXiv* abs/2401.14159 (2024). https://api.semanticscholar.org/CorpusID:267212047
Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. 2025. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 6121–6132.
Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.
Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhu, Jun Zhang, and Yikai Wang. 2025. 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*. 13695–13706.
Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, Kun Gai, Sainan Guo, Feng Han, Jingbin He, Kang He, et al. 2025. Kling-Omni Technical Report. *arXiv preprint arXiv:2512.16776* (2025).
Yuanpeng Tu, Hao Luo, Xi Chen, Sihui Ji, Xiang Bai, and Hengshuang Zhao. 2025. 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*. 1–11.
Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. 2018. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717* (2018).
Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314* (2025).
Chaoyang Wang, Peiye Zhuang, Tuan Duc Ngo, Willi Menapace, Aliaksandr Siarohin, Michael Vasilkovsky, Ivan Skorokhodov, Sergey Tulyakov, Peter Wonka, and Hsin-Ying Lee. 2025d. 4Real-Video: Learning generalizable photo-realistic 4D video diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 17723–17732.
Hanlin Wang, Hao Ouyang, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Qifeng Chen, Yujun Shen, and Limin Wang. 2025a. Levitor: 3d trajectory oriented image-to-video synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 12490–12500.
Hanlin Wang, Hao Ouyang, Qiuyu Wang, Yue Yu, Yihao Meng, Wen Wang, Ka Leong Cheng, Shuailei Ma, Qingyan Bai, Yixuan Li, et al. 2025b. The World is Your Canvas: Painting Promptable Events with Reference Images, Trajectories, and Text. *arXiv preprint arXiv:2512.16924* (2025).
Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. 2024c. Boximator: Generating rich and controllable motions for video synthesis. *arXiv preprint arXiv:2402.01566* (2024).
Qianqian Wang, Vickie Ye, Hang Gao, Weijia Zeng, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. 2025c. Shape of motion: 4d reconstruction from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9660–9672.
Wenguan Wang, Hongmei Song, Shuyang Zhao, Jianbing Shen, Sanyuan Zhao, Steven CH Hoi, and Haibin Ling. 2019. Learning unsupervised video object segmentation through visual attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3064–3074.
Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. 2024b. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems* 36 (2024).
Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.
Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. 2024a. Motionctrl: A unified and flexible motion controller for video generation. In *SIGGRAPH*.
Cong Wei, Quande Liu, Zixuan Ye, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhu Chen. 2025. Univideo: Unified understanding, generation, and editing for videos. *arXiv preprint arXiv:2510.08377* (2025).
Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023. 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*. 7623–7633.
Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T Barron, and Aleksander Holynski. 2025b. Cat4d: Create anything in 4d with multi-view video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 26057–26068.
Yuhui Wu, Liyi Chen, Ruibin Li, Shihao Wang, Chenxi Xie, and Lei Zhang. 2025a. Insvie-1m: Effective instruction-based video editing with elaborate dataset construction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 16692–16701.
Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. 2021. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9421–9431.
Yuxi Xiao, Jianyuan Wang, Nan Xue, Nikita Karaev, Yuri Makarov, Bingyi Kang, Xing Zhu, Hujun Bao, Yujun Shen, and Xiaowei Zhou. 2025. SpatialTrackerV2: 3D Point Tracking Made Easy. *ArXiv* abs/2507.12462 (2025).
Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. 2024. SpatialTracker: Tracking Any 2D Pixels in 3D Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20406–20417.
Shuzhou Yang, Xiaoyu Li, Xiaodong Cun, Guangzhi Wang, Lingen Li, Ying Shan, and Jian Zhang. 2025a. Gencompositor: generative video compositing with diffusion transformer. *arXiv preprint arXiv:2509.02460* (2025).
Xiangpeng Yang, Ji Xie, Yiyuan Yang, Yan Huang, Min Xu, and Qiang Wu. 2025b. Unified Video Editing with Temporal Reasoner. *arXiv preprint arXiv:2512.07469* (2025).
Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2024. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072* (2024).
Zixuan Ye, Xuanhua He, Quande Liu, Qiulin Wang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Qifeng Chen, and Wenhan Luo. 2025. UNIC: Unified In-Context Video Editing. *arXiv preprint arXiv:2506.04216* (2025).
Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. 2023. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089* (2023).
Meng You, Zhiyu Zhu, Hui Liu, and Junhui Hou. 2024. Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. *arXiv preprint arXiv:2405.15364* (2024).
Mark YU, Wenbo Hu, Jinbo Xing, and Ying Shan. 2025. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. *arXiv preprint arXiv:2503.05638* (2025).
Shoubin Yu, Jacob Zhiyuan Fang, Jian Zheng, Gunnar Sigurdsson, Vicente Ordonez, Robinson Piramuthu, and Mohit Bansal. 2024a. Zero-shot controllable image-to-video animation via motion decomposition. In *ACM MM*.
Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. 2024b. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048* (2024).
Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*. 3836–3847.
Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. 2025. Tora: Trajectory-oriented diffusion transformer for video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2063–2073.
Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. 2024. *Open-Sora: Democratizing Efficient Video Production for All*. https://github.com/hpcaitech/Open-Sora
Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. 2025. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint arXiv:2503.14489* (2025).
Bojia Zi, Weixuan Peng, Xianbiao Qi, Jianan Wang, Shihao Zhao, Rong Xiao, and Kam-Fai Wong. 2025. MiniMax-Remover: Taming Bad Noise Helps Video Object Removal. *ArXiv* abs/2505.24873 (2025). https://api.semanticscholar.org/CorpusID:279070405

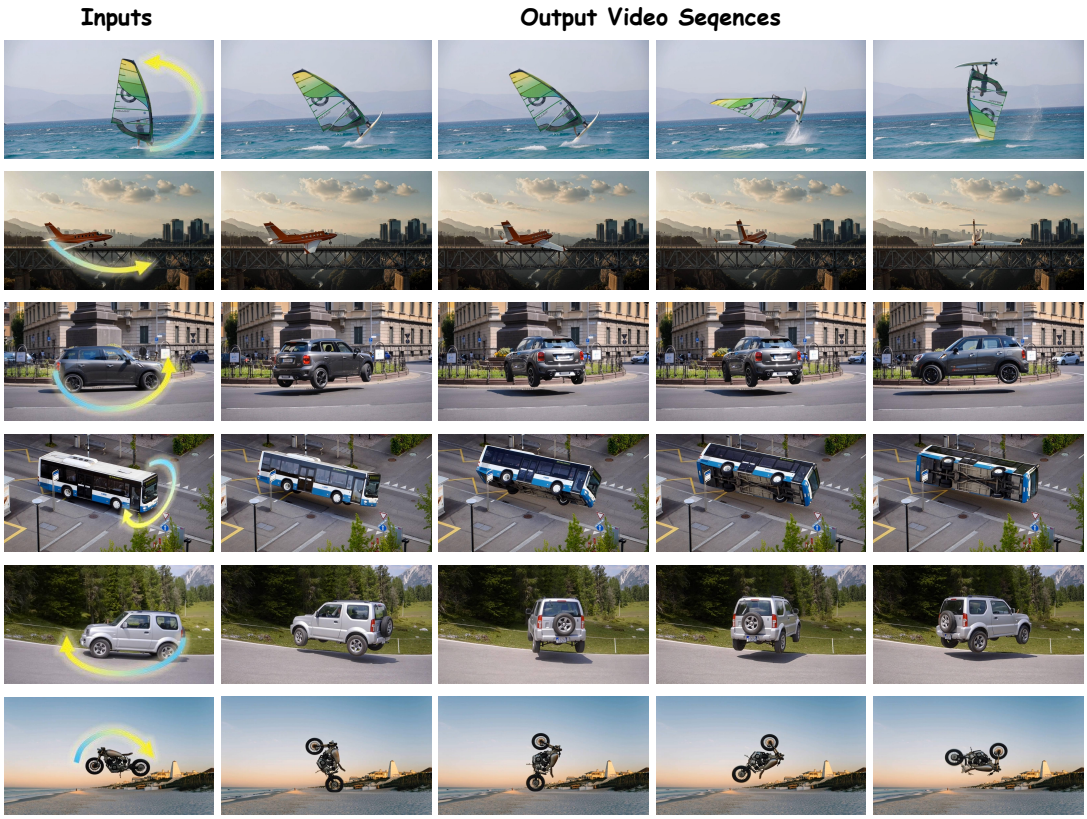

Fig. 9. Qualitative results of complex 3D rotations. Our framework enables the synthesis of challenging 3D motions, such as flips, rolls, and axial rotations. This process begins by reconstructing the target foreground assets using SAM3D to obtain a manageable 3D representation. We then apply precise 3D transformations to these reconstructed objects and extract the resulting trajectory as a motion guidance signal. By injecting this trajectory into our model, we can synthesize video sequences that maintain consistency during extreme pose changes.

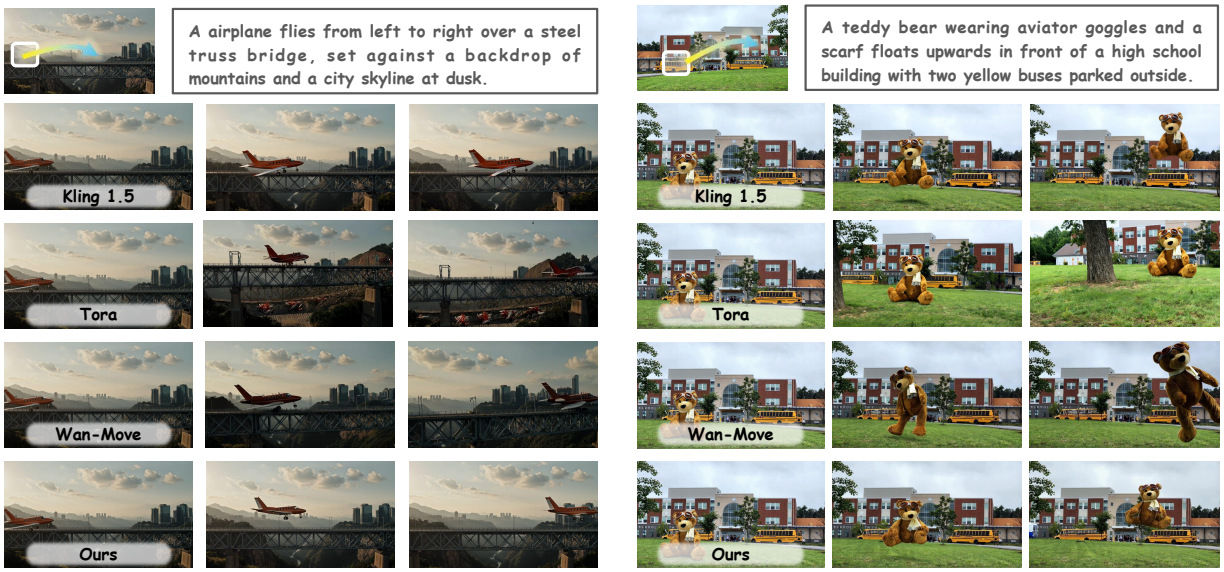

Fig. 10. Additional Qualitative Comparison on Static Foreground Compositing. We compare our method against trajectory-controlled I2V methods on two challenging scenarios. Our method demonstrates superior occlusion handling and trajectory adherence while maintaining visual fidelity.

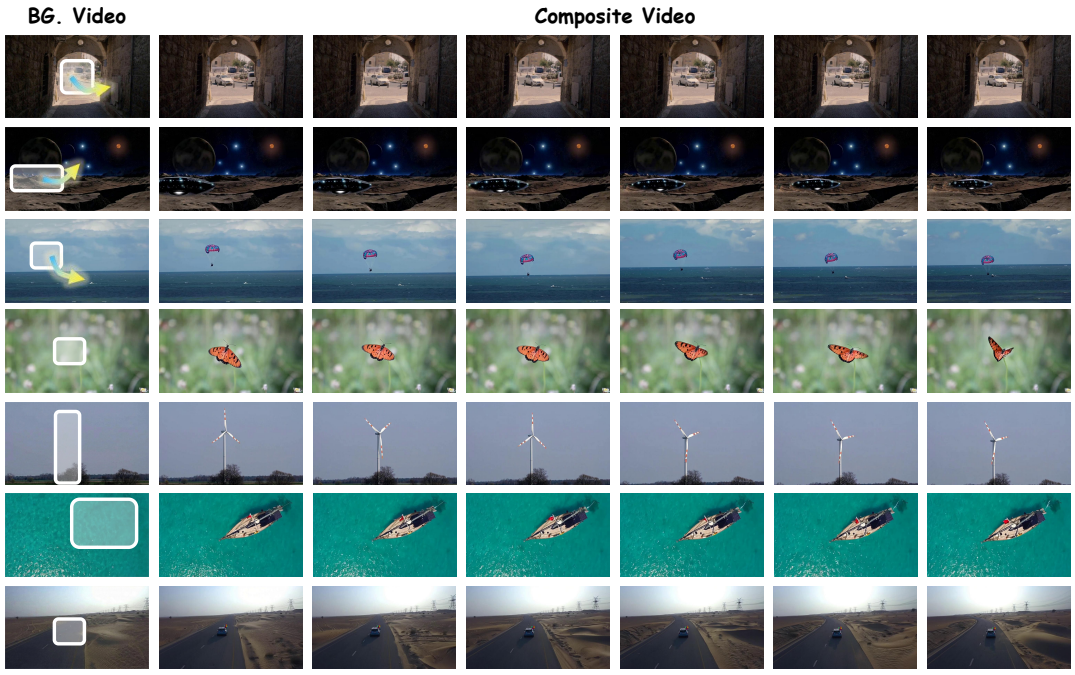

Fig. 11. Qualitative results of flexible video composition. The first column illustrates the Background Video annotated with user-defined spatial trajectories (indicated by white bounding boxes and motion arrows). The following columns display the Composite Video sequences generated by our model.

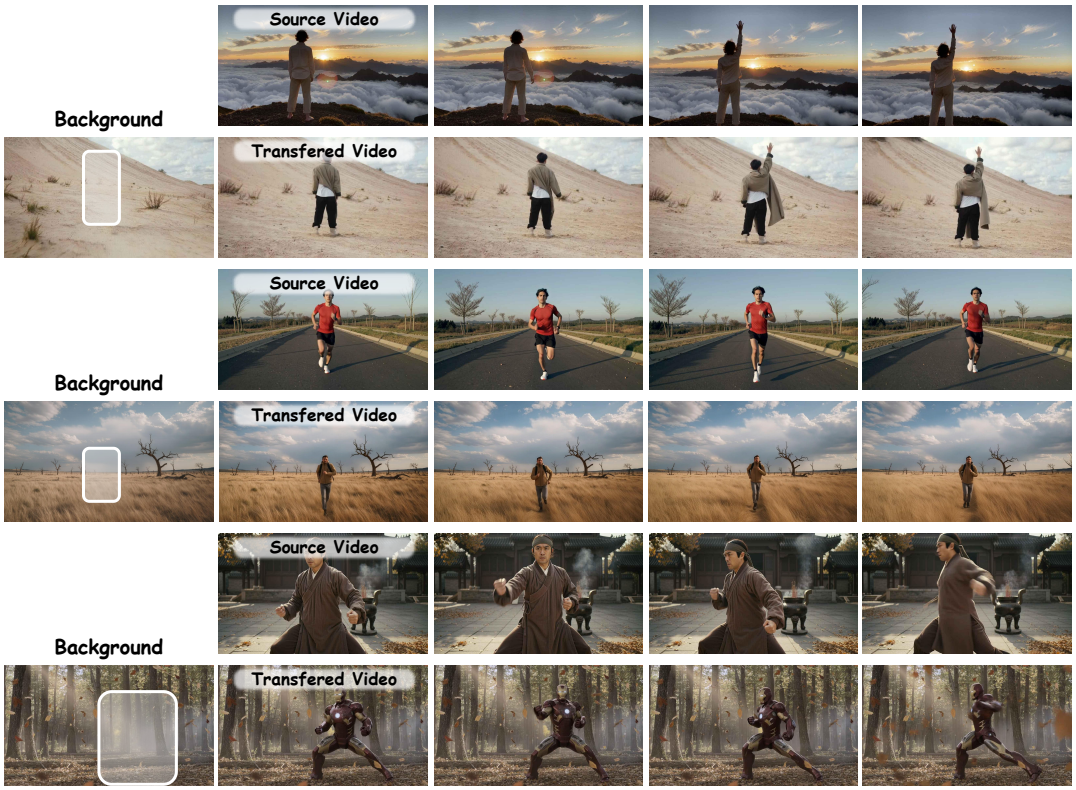

Fig. 12. Qualitative results of motion transfer. By mapping source trajectories to target characters via latent transport, our framework facilitates motion synthesis in new background. This approach supports non-pixel-aligned transfer, allowing for motion transfer across different character and backgrounds.