Tri-Prompting: Controllable Video Generation with Scene, Subject, and Motion Prompts

Zhenghong Zhou[1,2*] **Xiaohang Zhan**[1] **Zhiqin Chen**[1] **Soo Ye Kim**[1] **Nanxuan Zhao**[1]
Haitian Zheng[1] **Qing Liu**[1] **He Zhang**[1] **Zhe Lin**[1] **Yuqian Zhou**[1‡] **Jiebo Luo**[2‡]
[1]Adobe Research [2]University of Rochester [‡]Equal advising
Project Page

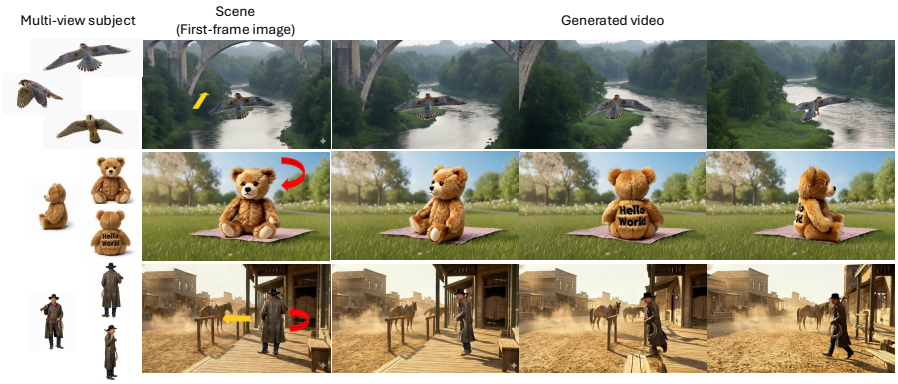


Figure 1: Tri-Prompting integrates scene control, multi-view subject control, and scene–subject motion control within one video diffusion model. Users can select a 3D-aware subject, insert or manipulate it in diverse scenes, and specify camera and subject motion through an offline control interface with keyboard, while maintaining appearance consistency with reference images. The figure shows scene/camera-only, subject-only, and joint scene-subject motion control. Yellow arrow → indicates camera trajectory, and Red arrow → indicates subject motion.

Abstract

Recent video diffusion models achieve strong visual quality and temporal coherence, but still lack coordinated control over scene layout, customized subject identity, and camera/subject motion. We study controllable video generation from three prompts: a first-frame scene image, 3D-aware multi-view subject references, and a motion-driving signal. This setting is challenging because background regions are often trackable under camera motion, while foreground subjects can rotate, self-occlude, and reveal new regions that require identity-consistent appearance. We introduce Tri-Prompting, a video diffusion framework that combines multi-view subject conditioning with dual-conditioned motion control. Tri-Prompting uses XYZ tracking points for visible background motion and low-resolution RGB proxies for foreground subject pose, allowing the model to recover fine appearance from multi-view references while retaining flexibility for plausible subject-scene interactions. An inference-time ControlNet scale schedule further balances motion controllability and visual realism. We evaluate Tri-Prompting against specialized baselines: DaS for motion reconstruction and Phantom for subject-driven generation. Tri-Prompting achieves competitive or improved reconstruction quality, stronger multi-view identity preservation, and better 3D consistency, while enabling 3D-aware subject insertion and in-scene manipulation.

*Work done during an internship at Adobe.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Recent video diffusion models deliver remarkable visual quality and temporal coherence [33, 46]. However, video creation is naturally organized around three questions: where the story happens, who appears in it, and how the camera and subject move. In this work, we instantiate these three factors as scene, subject, and motion prompts: a first-frame scene image, multi-view subject references, and a motion-driving signal. The goal is to generate a video that preserves the scene layout and 3D-aware subject identity while following the specified camera and subject motion. For example, starting from a single scene image, a creator may insert a new 3D-aware subject or manipulate an existing one, move the camera, and turn the subject around.

Table 1: Factorized controllability.

	Where Scene (first frame)	**Who** Multi-views of subject	**How** Cam. / Subj. motion
I2V (Wan) [33]	✓	✗	✗
Phantom [21]	✗	—	✗
DaS [12]	✓	—	—
Ours (Tri-Prompting)	✓	✓	✓

Notes. ✓ denotes full support, ✗ denotes no support, and — denotes partial or limited support. Specialized methods only handle isolated factors. *Phantom* generates videos from single-view subjects (—); *DaS* enables motion control video generation (—) but is limited to the first frame, and fails to maintain multi-view appearance consistency. In contrast, **Tri-Prompting** covers all three dimensions.

Existing methods cover only part of this tri-prompt setting. Motion-guided methods such as Motion-Prompting [11] and DaS [12] provide trajectory control, but rely mainly on first-frame appearance and struggle to preserve customized identity under large view changes. Subject-driven methods such as Phantom [21] preserve subject appearance from references, but lack camera and subject motion control. Thus, existing methods do not directly support first-frame scene control, multi-view subject identity, and motion control in one generation framework.

This setting poses two challenges. First, background and foreground motion have different trackability. Background regions are often trackable under camera motion, making tracking points effective for scene control. In contrast, foreground subjects can rotate, self-occlude, or deform, causing regions to disappear or become newly visible, which makes tracking-point controls unreliable for subject motion. Second, the subject should be represented as a complete object that can be viewed from different angles. A single reference image cannot reliably specify the side or back appearance, making multi-view references important for identity preservation under large viewpoint changes.

We introduce Tri-Prompting, a video diffusion framework for jointly controlling scene, subject, and motion. To address the different trackability of background and foreground motion, we use separate motion representations: XYZ trajectories for visible background surfaces and downsampled RGB grids for foreground subjects whose tracking points may disappear under rotation or self-occlusion. The low-resolution RGB proxy encodes 3D-aware subject pose without over-specifying fine appearance details, allowing the model to recover side and back appearance from multi-view references while retaining flexibility to synthesize realistic subject-scene interactions. We implement this design in two stages: the first stage learns to fuse the first-frame scene and multi-view subject references, and the second stage trains a ControlNet module for dual-conditioning motion control. At inference time, we further use a ControlNet scale schedule to reduce over-constraining from coarse motion cues and balance motion controllability with visual realism.

We evaluate Tri-Prompting through comparisons with the closest specialized baselines. For motion controllability, we compare with DaS on video reconstruction and show that dual conditioning improves reconstruction quality, especially in perceptual similarity. For multi-view subject-to-video generation, we compare with Phantom and observe consistent gains in video quality, multi-view identity similarity, and 3D consistency. Beyond these evaluations, we demonstrate 3D-aware subject insertion and in-scene manipulation workflows, where users can specify scene, subject, and motion prompts within the same framework. We fine-tune Tri-Prompting with 11k tuples ($<$7 hours of video) for fewer than 5k steps.

In summary, our contributions are as follows:

- **Tri-prompting for controllable video generation.** We propose Tri-Prompting, a single video diffusion framework that integrates scene, subject, and motion prompts for controllable video generation.
- **Dual-conditioning motion control with multi-view subject consistency.** We design a motion control module that combines XYZ trajectories for background motion and low-resolution RGB proxies for foreground subject motion, enabling decoupled scene/subject control while preserving identity with multi-view references.
- **3D-aware subject insertion / manipulation workflows and competitive results.** Tri-Prompting enables subject insertion and in-scene manipulation workflows, and achieves competitive or improved performance against the closest specialized baselines, including DaS for motion reconstruction and Phantom for subject identity preservation.

2 Related work

Text-to-Video and Image-to-Video Diffusion Models. Text-to-video (T2V) and image-to-video (I2V) diffusion models have progressed rapidly. Early video diffusion models [14, 39, 13, 2] extended text-to-image latent diffusion models by adding temporal layers to UNet backbones. Recent transformer-based video diffusion models [46, 4, 20, 22, 33, 26] represent videos as spatio-temporal tokens and use full attention to improve visual quality, temporal coherence, and motion diversity. Despite these advances, text or image prompts alone remain insufficient for precise control over scene layout, subject identity, and motion.

Motion-Guided Video Generation. Motion-guided video generation uses spatial or temporal cues such as pose, boxes, depth, camera trajectories, and tracking points [23, 44, 43, 35, 8, 30, 47, 12, 38, 11, 49]. Trajectory- and tracking-based methods are closely related to our motion prompt: MotionPrompting [11] uses 2D motion trajectories, while DaS [12] uses 3D tracking points to control visible scene motion. These methods are effective for moving visible content, but they do not address 3D-aware subject control: when a customized subject rotates or self-occludes, existing tracking points disappear, while newly visible surfaces have no corresponding tracks. Direct-a-Video [45] and MotionBooth [40] further support decoupled object and camera control, but mainly through 2D box controls rather than 3D-aware subject motion that reveals new views of a customized object. They also do not use a first-frame scene image for scene preservation. Frame In-N-Out [34] controls object entry and exit on an expanded canvas using 2D trajectories. CustomX [37] animates customized characters in a 3D Gaussian scene with text-guided actions and camera control. In contrast, Tri-Prompting uses a first-frame scene image, multi-view subject references, and separate background/foreground motion anchors to enable 3D-aware motion such as subject rotation.

Subject-Driven Video Customization. Subject-driven video customization aims to preserve the appearance of user-provided subjects across generated videos [17, 50, 9, 21, 5, 18]. Single-view references provide only partial appearance, making side and back views ambiguous under large viewpoint changes. Recent methods, including 3DreamBooth [19] and MV-S2V [31], use multi-view information to improve 3D subject consistency. Tri-Prompting combines multi-view subject conditioning with explicit camera and subject motion control to preserve identity under large view changes.

3 Method

3.1 Overview

Tri-Prompting consists of a video diffusion model guided by first-frame and multi-view reference images, and a motion control module driven by dual-conditioning 3D control cues, as shown in Fig. 2. This framework jointly controls scene, multi-view subject, and motion in video generation. It is also designed to preserve multi-view appearance identity of the given subject, and achieve flexible disentangled motion control between foreground object and background scenes.

During inference, Tri-Prompting requires three inputs: (1) an image $I \in \mathbb{R}^{H\times W\times 3}$ with a text prompt that defines the scene layout and initial composition; (2) multi-view images (up to three) $\{S_i\}$ of the same single subject, where each $S_i \in \mathbb{R}^{H\times W\times 3}$, which define subject identity and unseen appearance; (3) a motion-control video $M \in \mathbb{R}^{T\times H\times W\times C}$ containing XYZ points for background/camera motion

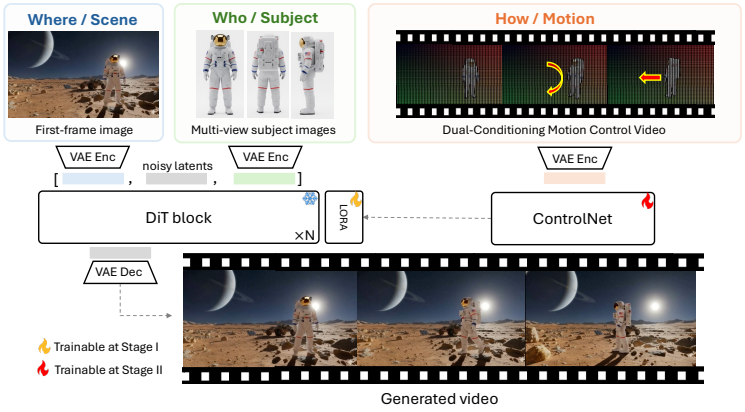


Figure 2: **Overview of the proposed Tri-Prompting.** Tri-Prompting integrates video diffusion with first-frame/multi-view images and a dual-conditioning motion anchor to jointly control scene, subject and motion. We use two-stage training: first optimizing LoRA for scene and subject control, followed by ControlNet finetuning for motion control. The framework preserves multi-view identity while supporting decoupled motion control between the foreground and background.

and low-resolution RGB proxies for foreground subject motion. This video can be generated from a reference video, pose-matrix transformations, or an offline control interface with keyboard. The output is a video $V \in \mathbb{R}^{T\times H\times W\times C}$.

Sec. 3.2 introduces the proposed multi-view subject I2V model, and Sec. 3.3 presents the ControlNet-based motion control module with dual 3D cues. Finally, Sec. 3.4 describes the inference pipeline and different practical workflows.

3.2 Multi-view Subject-Image-to-Video

Tri-Prompting requires two stages of training to make the optimization more stable, effective, and trackable. Stage 1 focuses on extending the single-view subject-to-video model into a joint image and multi-view subject-to-video model, establishing foundational control over both the background scene (via the first-frame image) and 3D-aware subject identity (via multi-view reference images). Specifically, we encode the first-frame image and multi-view subject images with the base video diffusion's VAE encoder to get $z_I \in \mathbb{R}^{1\times \frac{H}{sc}\times \frac{W}{sc}\times C}$ and $z_S \in \mathbb{R}^{k\times \frac{H}{sc}\times \frac{W}{sc}\times C}$, where k is the number of multi-view subject reference images, sc is the spatial compression factor, and C is the latent channel number. We then prepend z_I to the original noisy video latents $z_V \in \mathbb{R}^{\frac{T}{tc}\times \frac{H}{sc}\times \frac{W}{sc}\times C}$, where tc is the temporal compression ratio of the VAE, and subsequently append z_S to be the input token sequence z_{seq} of DiT blocks:

$$z_{seq} \leftarrow [z_I, z_V, z_S] \in \mathbb{R}^{(1+\frac{T}{tc}+k)\times \frac{H}{sc}\times \frac{W}{sc}\times C} \tag{1}$$

The order of the three latent components can be swapped if full self-attention is used in the video diffusion model. We apply LoRA [15] on the attention and MLP blocks to condition on z_I for first-frame generation and on z_S for cross-view identity. During decoding, only z_V is kept to generate the final video V.

3.3 Dual-Conditioning Motion Control Module

The first stage already yields a video diffusion model that follows the first-frame scene and preserves identity across multiple views of the subject. In the second stage, we add explicit motion control over both the scene and the subject with dual control cues.

We first revisit two types of explicit control signals for motion generation: 3D tracking points and 2D RGB. 3D tracking points (XYZ trajectories) are effective control signals in previous works, such as DaS [12], but they mainly constrain visible and trackable surfaces. Because XYZ tracks are recovered from single-view videos, they capture only the geometry directly visible to the camera. For foreground subjects, rotation and self-occlusion can reveal newly visible regions that have no

corresponding tracks, making tracking-point-only control unreliable for large viewpoint changes. Another potential control condition is a dense RGB pixel signal produced by optical flow, as in EPiC [38]. Dense RGB pixels tend to encourage the model to learn hole-filling behaviors like inpainting, yielding artifacts around hole regions.

To address these problems, we define a dual-conditioning signal: 3D tracking points for background and downsampled RGB proxies for foreground. Intuitively, XYZ tracking points are suitable for visible and trackable background surfaces under camera motion, while low-resolution RGB proxies encode coarse 3D-aware subject pose and spatial motion for foreground. These proxies intentionally avoid specifying fine appearance details, encouraging the model to rely on multi-view references and its generative prior to recover subject appearance and plausible subject-scene interactions. Root-6DoF controls [10] require accurate object poses and canonical frames, which are difficult to obtain for diverse open-world subjects. Low-resolution RGB proxies can be constructed directly from videos or renders without object-specific pose annotations.

Specifically, we define the anchor motion control video as $M = \{M_{scene}, M_{subject}\}$. For scene (background) control, we follow DaS to construct an XYZ tracking point video $M_{scene} \in \mathbb{R}^{T\times H\times W\times C}$. Each point's 3D coordinates (XYZ) are determined based on its position and depth in the first frame. These coordinates are then normalized to [0, 1] and converted to pseudo-RGB c_i. The color of the same tracking point does not change to keep the identity of the point. For subject (foreground) control, we use a low-resolution downsampled RGB point proxy $M_{subject} \in \mathbb{R}^{T\times H\times W\times C}$ obtained by downsampling the subject pixels into a fixed grid (e.g., 70×70) within the subject region. We composite these two conditions into a single anchor video in a spatially exclusive manner. This proxy provides sufficient cues of the camera pose as DaS [12], and provides stronger guidance on the object motion. Missing appearance details are complemented by the multi-view subject references via self-attention layers and by the base model prior.

We leverage this proxy and develop a ControlNet on top of the stage 1 base model. We copy the trained weights from the stage 1 model and add zero-initialized layers following the ControlNet architecture. The same video diffusion VAE encoder encodes M to $z_M \in \mathbb{R}^{\frac{T}{tc}\times\frac{H}{sc}\times\frac{W}{sc}\times C}$. Similarly to Sec. 3.2, the input of the ControlNet DiT block is generated by concatenating z_I, z_M and z_S. And only the z_M of output will be used to update z_V.

$$z_V \leftarrow z_V + s \cdot \mathrm{ControlNet}([z_I, z_M, z_S])[1 : 1 + \frac{T}{tc}] \tag{2}$$

where s represents the guidance scale of ControlNet. During the stage 2 finetuning, the weights of the base model are fixed, and only the ControlNet's weights are updated.

3.4 Inference

3.4.1 Inference-time ControlNet Scale Schedule

During training, the low-resolution RGB condition is derived directly from ground truth video, ensuring perfect alignment with the target. At inference time, user-specified 3D transformations provide coarse motion but may not contain natural micro-motions, such as leg lifts during walking, leading to rigid results. Although diffusion priors can bring realism, they remain limited when over-constrained by these motion cues.

To address this issue, we introduce an optional ControlNet scale schedule to balance the trade-off between controllability and video realism. While the model is trained with a fixed ControlNet scale of 1.0, during inference we gradually reduce this influence to prevent over-constraining the generation. Specifically, for a sampling process with 50 denoising steps, we linearly anneal the scale during the first N_{decay} steps and keep it constant thereafter:

$$\mathrm{s}(t) = \begin{cases} 1 - \frac{t}{N_{\text{decay}}}(1 - s_{\min}) & \text{if } t \le N_{\text{decay}}, \\ s_{\min} & \text{otherwise.} \end{cases} \tag{3}$$

where $s_{\min}$ is the final ControlNet scale after decay.

3.4.2 Inference Workflow

Tri-Prompting enables several new workflows that were previously difficult or impossible. We introduce two novel workflows using an offline control interface and an inference pipeline that support identity-preserved motion control. The base model accepts text prompts for further flexibility.

Users can choose the scene (first frame image), select the character (multi-view subject), and drive the scene and subject motion. We developed an offline control interface that allows users to drive motion via keyboard controls for 3D subject translation/rotation and camera pose. Given subject reference images, we reconstruct a 3D subject using the Gaussian output of TRELLIS [41], render downsampled 2D projections per frame, and paste it as the foreground RGB guidance. For the background, we estimate the first-frame depth with DepthPro [3] and convert camera pose into XYZ point trajectories. These two spatially exclusive signals form the dual conditioning for motion control.

Insertion of 3D Subjects into Scenes and Joint Control. Given a background scene image and a 3D subject, we first create a harmonized initial frame by inserting the subject's initial 2D projection using an image editing model (e.g., Gemini [32], FluxKontext [1], or Photoshop Generative Fill). We then provide three representative subject views to the model. Starting from this frame, users can control camera and subject motion independently or jointly: (i) camera control is specified by transforming only the background XYZ points; (ii) subject control is specified by applying translations/rotations to the reconstructed 3D subject. The low-resolution foreground guidance steers motion rather than appearance, while multi-view references preserve identity and 3D-consistent details.

Manipulation of 3D Subjects in a Scene and Joint Control. Given a scene image containing multiple subjects, we use it as the first frame, obtain a target mask via SAM 2 [28], and reconstruct the selected subject in 3D with SAM 3D [7]. We render representative multi-view references from the reconstructed 3D asset, optionally refined by an image editing model. With the same dual-conditioning construction, users can control camera pose and subject motion to manipulate the scene with identity-preserved, motion-controlled generation.

4 Experiment

4.1 Training Data

Training Tri-Prompting requires three main assets from videos (V): the first frame (I), three multi-view reference images ($\{S_i\}$) of the main subject in the scene, and a synthesized motion anchor video (M). To make the training efficient, we require the dataset videos to contain a single salient subject with extreme pose changes. The categories of the subjects better cover both rigid objects and non-rigid characters. Therefore, we constructed 11k tuples $(I, \{S_i\}, M, V)$ (total $<$7 hour-video), comprising 9.7k game videos from the OmniWorld-Game dataset [48] and 1.3k real-world videos from the CO3D dataset [29]. Despite this focused training, we observe qualitative generalization across scenes/subjects/poses and diverse styles (e.g., anime, film); see Sec. 4.4 for illustration.

Each tuple was constructed as follows: (1) First frame (I): We took the first frame of each video as the scene image. (2) Multi-view subject images ($\{S_i\}$): For OmniWorld-Game, we manually identified and cropped three views of the same subject across frames. For the CO3D dataset, which is object-centric with rotating views, we directly sampled frames around the object as the multi-view subject references $\{S_i\}$. (3) Motion control videos (M). Each dataset provides the masks of the salient subjects. We separate the background and foreground subject to build the dual-conditioning motion control videos. For the background, we follow DaS to extract an XYZ tracking video: 3D points are tracked by SpatialTracker [42] across frames and normalized. For the subject, we construct low-resolution RGB points by downsampling the whole video from 832×480 to a fixed grid (70×70), and keeping only grids within the subject mask.

4.2 Experiment Details and Settings

We use a two-stage training strategy in Tri-Prompting. In stage 1, we fine-tune the Phantom S2V 14B backbone [21] with LoRA (rank 64) on attention and MLP blocks. Training utilizes AdamW (lr 1×10^{-4}), with a batch size of 8 for 2500 steps. This stage takes 20 hours on 8 A100 GPUs. In stage 2, we initialize a ControlNet by copying the first 18 layers from the first stage checkpoint. Then we freeze the base diffusion model and finetune the added ControlNet modules. Control injections are

Table 2: **Comparison with specialized baselines.** (a) Video reconstruction on DAVIS. (b) Video quality, multi-view identity, and 3D consistency on our evaluation set; geometry scores use 84 videos successfully reconstructed across all settings.

(a) Comparison with DaS on DAVIS

	Motion control	**Base Model**	**PSNR↑**	**SSIM↑**	**LPIPS↓**
DaS [12]	XYZ tracking	CogVideoX (5B) [46]	*16.4916*	**0.4123**	0.2725
Ours	XYZ tracking	Phantom (14B) [21]	15.9306	0.3708	0.2786
Ours	XYZ tracking & RGB	Phantom (14B) [21]	**16.5130**	*0.4017*	**0.2395**

(b) Comparison with Phantom and OmniVCus on VBench and Multi-view Consistency

Method	Motion Anchor	VBench Quality ↑				MV Subj. ID ↑		3D Consis. ↓
		Aes.	Mot.	Dyn.	Img.	$\mathcal{I}_{\text{dino}}$	$\mathcal{I}_{\text{clip}}$	E_{align}
Phantom (3-view)	No	0.655	0.988	0.742	0.709	0.723	0.855	0.034
OmniVCus (3-view montage)	Mask	0.618	0.987	0.613	0.703	0.703	0.846	0.044
Ours stage1 (3-view)	No	0.660	0.987	0.682	0.705	0.732	0.861	0.027
Ours stage2 (3-view)	Yes	**0.665**	**0.989**	**0.903**	**0.715**	**0.746**	**0.862**	**0.025**

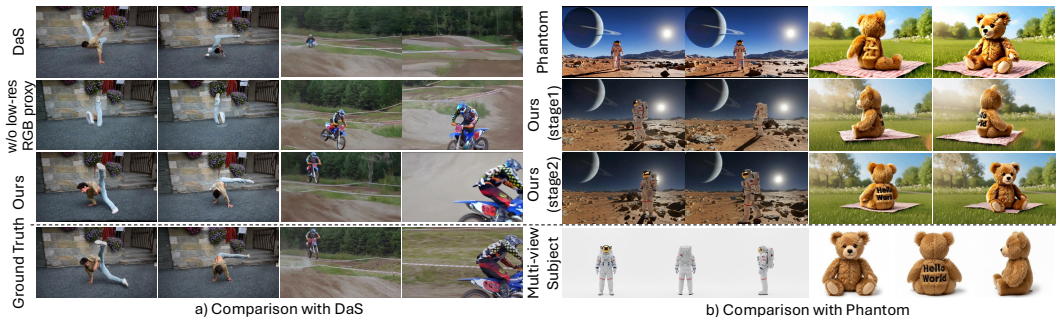


Figure 3: **Comparison with DaS and Phantom.** Tri-Prompting improves motion robustness over tracking-only control and preserves multi-view subject identity under large pose changes.

zero-initialized (zero-conv). Training also utilizes AdamW (lr 1×10^{-4}), with a batch size of 32 for 2074 steps. The second stage takes 28 hours using 32 A100 GPUs.

4.3 Comparison with Baselines

We compare with DaS [12] for video reconstruction, and Phantom [21] and OmniVCus [5] for subject identity, video quality, and 3D consistency. Additional comparisons with MotionBooth are provided in Sec. D.3.

Comparison with DaS: Following DaS, we evaluate the video reconstruction performance under multi-conditioning using PSNR, SSIM, and LPIPS, demonstrating motion controllability of our model. We evaluate on 50 videos from the DAVIS dataset [25], covering diverse motion patterns, scenes, and subjects. We use the official DaS checkpoint to test on DAVIS, as shown in Tab. 2(a).

DaS reconstructs the video from the first frame and its XYZ tracking points. Tri-Prompting additionally uses low-resolution foreground RGB and multi-view subject references, yielding higher PSNR and lower LPIPS. Its lower SSIM and weaker XYZ-only result may reflect Phantom's focus on reference-conditioned identity compared with CogVideoX's first-frame I2V focus. We use Phantom for its subject-reference interface and isolate the RGB contribution through same-backbone ablations (Sec. 4.5).

As illustrated in Fig. 3, our method maintains robust motion alignment under extreme human actions where tracking-point-only methods fail and hallucinate as points disappear. Crucially, multi-view references enable the accurate recovery of intricate details like text and patterns that are occluded in the first frame—details that baseline models must otherwise hallucinate.

Comparison with Phantom: We compare against Phantom across three dimensions: video quality, multi-view subject similarity, and 3D consistency. We define 106 test cases from scene, subject, and

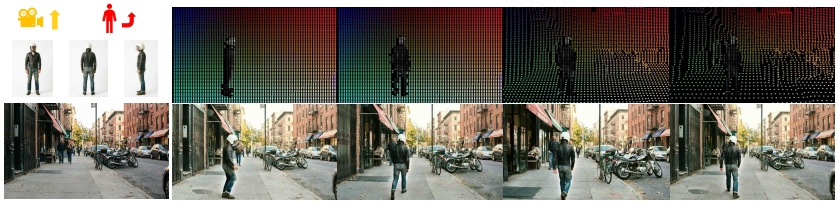


Figure 4: **Applications.** Tri-Prompting supports 3D-aware subject insertion and in-scene manipulation with camera/subject motion control. The multi-view identity is also preserved, providing strong 3D consistency. Please see more video results in appendix.

motion inputs before generation and evaluate all 318 outputs (3 random seeds). For each case, both models share the scene, 3-view subject images, and prompt; ours can also take additional motion anchors.

For video quality, we use four metrics from VBench [16]: aesthetic quality, dynamic degree, imaging quality, and motion smoothness. For multi-view ID, $\mathcal{I}_{dino/clip} = \text{mean}_i \max_j \text{sim}(\mathbf{f}_i, \mathbf{s}_j)$ compares DINO [24]/CLIP [27] features of segmented subject regions [28] in generated frames $\mathbf{f}$ with subject images $\mathbf{s}$. For 3D consistency, we reconstruct 3D points $\mathcal{P}_g$ and $\mathcal{P}_s$ from subject region of generated videos and subject images using reconstruction model π^3 [36] and measure 3D alignment error $E_{\text{align}} = \text{mean}_{\mathbf{x}\in\mathcal{P}_g} \min_{\mathbf{y}\in\mathcal{P}_s} \|\mathbf{x} - \mathbf{y}\|_2$. In Tab. 2(b), we evaluate rigid subjects (*e.g.*, cars, chairs, and bags) on 84 outputs successfully reconstructed across all evaluated settings; protocol details are in Sec. D.1. Results covering all subjects are reported in Sec. E.

As shown in Tab. 2(b), Tri-Prompting outperforms Phantom across all metrics, reducing E_{align} from 0.034 to 0.025. This result underscores a practical advantage: while Phantom mainly enforces 2D appearance consistency, our method better preserves 3D integrity across frames while following motion control. Qualitative results in Fig. 3 further highlight these differences; Phantom exhibits structural distortions during multi-view fusion—for instance, an astronaut's body may face backward while moving forward, and a rotating teddy bear suffers from warping and loss of shape. In contrast, Tri-Prompting maintains more consistent identity and 3D structure under such large pose changes.

Comparison with OmniVCus. We evaluate the released OmniVCus checkpoint on the same 106 cases and three seeds, using text, foreground-mask videos, and a fixed montage of the three subject views. Tri-Prompting improves video quality, multi-view identity, and 3D consistency over OmniVCus (Tab. 2(b)). Single-reference results and input details are provided in Sec. D.1.

4.4 Applications

Some qualitative results are shown in Fig. 4, highlighting both (i) inserting multi-view subjects into a scene and (ii) manipulating existing subjects within a scene. In both applications, the resulting foreground-background interactions remain visually coherent. For instance, under camera-pose changes, the cowboy and Wukong (monkey) exhibit plausible non-rigid motion (e.g., walking) consistent with the scene dynamics. Additionally, the multi-view appearance from the reference subject images is preserved, providing strong identity consistency.

Tri-Prompting provides robust, generalizable control over scene, subject, and motion. Although fine-tuned on a mix of game [48] and common subjects videos [29], we observe generalization to scenes/subjects/motion and diverse styles (anime/movie/real-world) in Fig. 4 and Supp., demonstrating our model works well on out-of-distribution samples.

4.5 Ablation Studies

We ablate dual-conditioning, training stages, reference views, ControlNet scale scheduling, and RGB resolution. Public 360Motion results in Tab. 3 are analyzed in Sec. 4.6.

Dual-conditioning vs Tracking Points Only. To evaluate the benefits of RGB conditioning, we compare different motion control signals on the same Phantom base model, as shown in the second and third rows of Tab. 2(a). Dual-conditioning achieves superior PSNR, SSIM, and LPIPS compared to using XYZ tracking alone. This demonstrates that incorporating RGB proxies provides more reliable guidance for extreme poses and 3D motion. In contrast, XYZ tracking alone often fails to propagate complex dynamics due to sparse or missing points.

Effect of Training Stages. We observe progressive performance gains across our two-stage training strategy (Tab. 2(b)). While Stage 1 establishes a robust foundation for multi-view identity, the introduction of dual-conditioning motion control in Stage 2 serves as more than a functional addition—it acts as a guiding anchor that further bolsters overall video quality, multi-view identity, and 3D consistency. This progression suggests that identity-focused training and motion-guided refinement work together to improve appearance preservation and motion control.

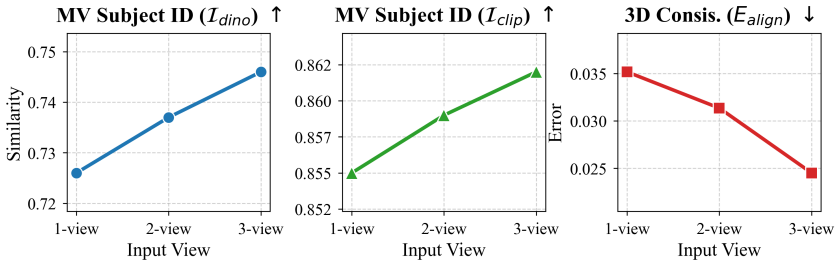


Figure 5: **Effect of multi-view subject images.** More views improve multi-view ID and 3D consistency, reducing ambiguity during pose changes.

Effect of Multi-view Subject Images. We finetune and infer models with 1/2/3 views to test the effects of multi-view subject views. As shown in Fig. 5, all metrics improve with more views. Providing 3 views reduces the 3D alignment error (E_{align}) from 0.035 to 0.025 and consistently boosts identity similarity. Single-reference input is supported, while additional views reduce appearance ambiguity during pose changes.

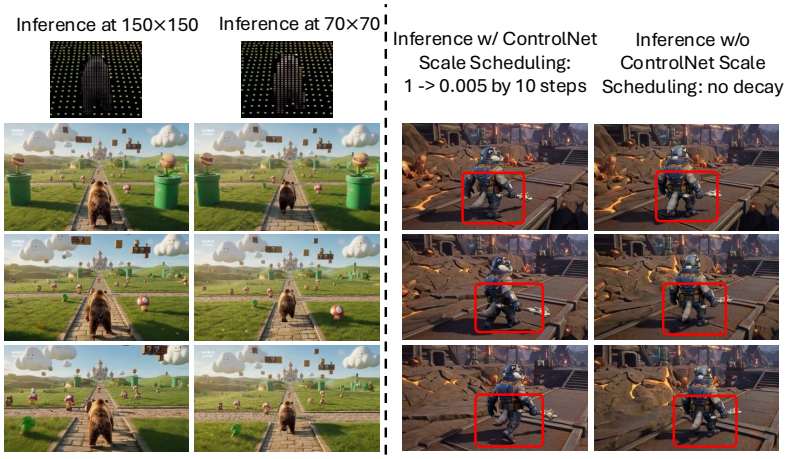


Figure 6: **Ablation study.** Left: Low-resolution RGB enables plausible subject-scene interactions. Right: ControlNet scale decay reduces rigid motion.

Effect of ControlNet Scale Scheduling. ControlNet scale scheduling balances controllability and visual realism. We compare constant scale 1.0 with decay from 1.0 to 0.005 over the first 10 of 50 denoising steps, using the same settings for all videos on both datasets. In Fig. 6, decay relaxes the

rigid translation cue and enables a more natural leg-lift gait, but may reduce motion controllability. On both datasets, decay improves multi-view identity with lower Bbox/Mask IoU (Tab. 3).

Low-res RGB vs. High-res RGB. As shown in Fig. 6, the low-resolution RGB condition provides greater flexibility for plausible subject-scene interactions, while higher-resolution RGB better preserves the rigid properties of objects when needed. Users can also adjust the RGB resolution during inference to control the degree of rigidity in the object's motion.

4.6 Generalization Studies

Our 106-case evaluation set contains no training cases and tests Tri-Prompting's out-of-distribution (OOD) generalization to unseen scene/subject/pose combinations across anime, movie, game, and real-world styles. We further evaluate on the independently constructed public 360Motion benchmark [10] (35 single-entity cases from the held-out C12 split, three seeds per case). 360Motion is in-domain for 3DTrajMaster but unseen during our training. Methods use their native conditioning inputs, as detailed in Sec. D.2. On 360Motion, constant scale yields higher Bbox/Mask IoU than 3DTrajMaster (0.632/0.510 vs. 0.590/0.435), while scale decay improves DINO identity similarity over Phantom (0.654 vs. 0.639) with comparable CLIP similarity (Tab. 3). This external evaluation further supports generalization of motion control and identity preservation.

Table 3: **Generalization and effects of ControlNet scale scheduling.** Motion adherence and identity on our evaluation set (318 outputs) and public 360Motion (105 outputs). Scale decay improves identity while reducing motion adherence.

Method	Our evaluation set				Public 360Motion			
	Bbox IoU↑	Mask IoU↑	$\mathcal{I}_{\text{dino}}$ ↑	$\mathcal{I}_{\text{clip}}$ ↑	Bbox IoU↑	Mask IoU↑	$\mathcal{I}_{\text{dino}}$ ↑	$\mathcal{I}_{\text{clip}}$ ↑
Phantom	—	—	0.723	0.855	—	—	0.639	0.874
3DTrajMaster	0.652	0.540	—	—	0.590	0.435	—	—
Ours, constant scale	**0.861**	**0.790**	0.733	0.858	**0.632**	**0.510**	0.627	0.864
Ours, scale decay	0.775	0.684	**0.746**	**0.862**	0.535	0.411	**0.654**	**0.875**

4.7 Limitation

Highly symmetric multi-view references can occasionally cause brief identity inconsistencies, such as transient flips, which can be alleviated by asymmetric references or stronger motion constraints. The pipeline also depends on external segmentation, reconstruction, and view-generation components; severe errors in their outputs can degrade generation. Large camera rotations can weaken background control when the XYZ points visible in the first frame leave the view. Our 14B model targets offline generation, taking about 3.85 minutes on eight A100-80GB GPUs for a 49-frame, 832×480 video with 50 denoising steps (Sec. C.1). Video diffusion acceleration could reduce this latency. The current method controls one primary subject and does not support complex multi-subject control.

5 Conclusion

We present Tri-Prompting, a video diffusion framework that integrates scene, subject, and motion prompts. By combining multi-view references with a dual-conditioning module—XYZ coordinates for backgrounds and low-resolution RGB proxies for foregrounds—we recover detailed, 3D-consistent subjects from coarse motion cues. In evaluations, Tri-Prompting achieves competitive or improved performance against specialized baselines: it obtains competitive PSNR/SSIM and improved perceptual similarity compared to DaS, and outperforms Phantom in multi-view subject-to-video generation. To support representative applications such as insertion and manipulation of 3D-aware subjects in a scene, we further develop a control interface to construct the motion video. Tri-Prompting supports scene/subject motion control across diverse scenes and styles. In the future, we plan to explore more efficient designs to reduce inference time, moving toward an interactive experience that can better support world-simulator-style applications.

Acknowledgments

This work was partially supported by the university's Goergen Institute for Data Science and Artificial Intelligence.

References

[1] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025.

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

[3] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *ICLR*, 2025.

[4] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

[5] Yuanhao Cai, He Zhang, Xi Chen, Jinbo Xing, Yiwei Hu, Yuqian Zhou, Kai Zhang, Zhifei Zhang, Soo Ye Kim, Tianyu Wang, et al. Omnivcus: Feedforward subject-driven video customization with multimodal control conditions. *arXiv preprint arXiv:2506.23361*, 2025.

[6] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

[7] Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025.

[8] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In *AAAI*, volume 39, pages 2403–2410, 2025.

[9] Zhengcong Fei, Debang Li, Di Qiu, Jiahua Wang, Yikun Dou, Rui Wang, Jingtao Xu, Mingyuan Fan, Guibin Chen, Yang Li, et al. Skyreels-a2: Compose anything in video diffusion transformers. *arXiv preprint arXiv:2504.02436*, 2025.

[10] Xiao Fu, Xian Liu, Xintao Wang, Sida Peng, Menghan Xia, Xiaoyu Shi, Ziyang Yuan, Pengfei Wan, Di Zhang, and Dahua Lin. 3DTrajMaster: Mastering 3D trajectory for multi-entity motion in video generation. *arXiv preprint arXiv:2412.07759*, 2024.

[11] Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, Chen Sun, et al. Motion prompting: Controlling video generation with motion trajectories. In *CVPR*, 2025.

[12] Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In *SIGGRAPH*, 2025.

[13] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[14] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022.

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

[16] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024.

[17] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *CVPR*, pages 6689–6700, 2024.

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

[19] Hyun-kyu Ko, Jihyeon Park, Younghyun Kim, Dongheok Park, and Eunbyung Park. 3dreambooth: High-fidelity 3d subject-driven video generation model. *arXiv preprint arXiv:2603.18524*, 2026.

[20] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[21] Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. *ICCV*, 2025.

[22] Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, et al. Step-video-t2v technical report: The practice, challenges, and future of video foundation model. *arXiv preprint arXiv:2502.10248*, 2025.

[23] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *AAAI*, volume 38, pages 4117–4125, 2024.

[24] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[25] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, pages 724–732, 2016.

[26] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[28] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[29] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *ICCV*, 2021.

[30] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *SIGGRAPH*, pages 1–11, 2024.

[31] Ziyang Song, Xinyu Gong, Bangya Liu, and Zelin Zhao. Mv-s2v: Multi-view subject-consistent video generation. *arXiv preprint arXiv:2601.17756*, 2026.

[32] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[33] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[34] Boyang Wang, Xuweiyi Chen, Matheus Gadelha, and Zezhou Cheng. Frame In-N-Out: Unbounded controllable image-to-video generation. In *Advances in Neural Information Processing Systems*, volume 38, pages 3291–3324, 2025.

[35] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *ECCV*, pages 55–72. Springer, 2024.

[36] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025.

[37] Yitong Wang, Fangyun Wei, Hongyang Zhang, Bo Dai, and Yan Lu. Customx: Unified character, action, and scene customization in video world models. *arXiv preprint arXiv:2512.17796*, 2025.

[38] Zun Wang, Jaemin Cho, Jialu Li, Han Lin, Jaehong Yoon, Yue Zhang, and Mohit Bansal. Epic: Efficient video camera control learning with precise anchor-video guidance. *arXiv preprint arXiv:2505.21876*, 2025.

[39] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, pages 7623–7633, 2023.

[40] Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation. *NeurIPS*, 37:34322–34348, 2024.

[41] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024.

[42] Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. Spatialtracker: Tracking any 2d pixels in 3d space. In *CVPR*, 2024.

[43] Jinbo Xing, Hanyuan Liu, Menghan Xia, Yong Zhang, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Tooncrafter: Generative cartoon interpolation. *TOG*, 43(6):1–11, 2024.

[44] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, et al. Make-your-video: Customized video generation using textual and structural guidance. *IEEE Transactions on Visualization and Computer Graphics*, 31(2):1526–1541, 2024.

[45] Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *SIGGRAPH*, pages 1–12, 2024.

[46] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *ICLR*, 2025.

[47] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024.

[48] Yang Zhou, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Haoyu Guo, Zizun Li, Kaijing Ma, Xinyue Li, Yating Wang, et al. Omniworld: A multi-domain and multi-modal dataset for 4d world modeling. *arXiv preprint arXiv:2509.12201*, 2025.

[49] Zhenghong Zhou, Jie An, and Jiebo Luo. Latent-reframe: Enabling camera control for video diffusion models without training. In *ICCV*, pages 12779–12789, 2025.

[50] Shaobin Zhuang, Kunchang Li, Xinyuan Chen, Yaohui Wang, Ziwei Liu, Yu Qiao, and Yali Wang. Vlogger: Make your dream a vlog. In *CVPR*, pages 8806–8817, 2024.

In the supplementary material, we provide additional qualitative results (Sec. A), details of the offline motion-control interface (Sec. B), implementation details and inference cost (Sec. C), additional baseline comparisons (Sec. D), 3D consistency analysis (Sec. E), and failure-case analysis (Sec. F).

A More Visualization Results

We provide more results of our method to illustrate its performance. These results include cases from games, movies, and real-world cases. **We strongly suggest checking the attached index.html file for the video results**.

In Fig. 9, we use a Cowboy video as the base to demonstrate representative controls across scene, subject, and motion. When the background scene is changed to a Times Square /old street and the subject is replaced by characters like a bear or anime character, the results successfully maintain the multi-view appearance of the subject and harmonious subject-scene interaction effects, including shadow casting and smooth motion (e.g., movement and turning). The final two rows illustrate object motion control (a 360° in-place rotation) and camera control (moving forward). These results demonstrate the robustness of our method, confirming its ability to adhere to all input conditions (scene image, multi-view subject, and motion) and to generate high-quality video with realistic interaction effects across them.

Fig. 10 presents subject insertion applications in real-world cases, covering humans, animals, and vehicles. The results show that our method remains effective beyond stylized domains, producing high-quality inserted subjects that are consistent with the input appearance and motion while blending naturally into real scenes.

Fig. 11 focuses on subject manipulation applications and shows that the method generalizes well across diverse fictional domains, including anime, movie, and game scenes. The generated videos align with the given subject images, preserving details (e.g., the horse and man during turnarounds). They also track the input motion while generating realistic interaction effects in various environments.

B Offline Motion-Control Video Construction

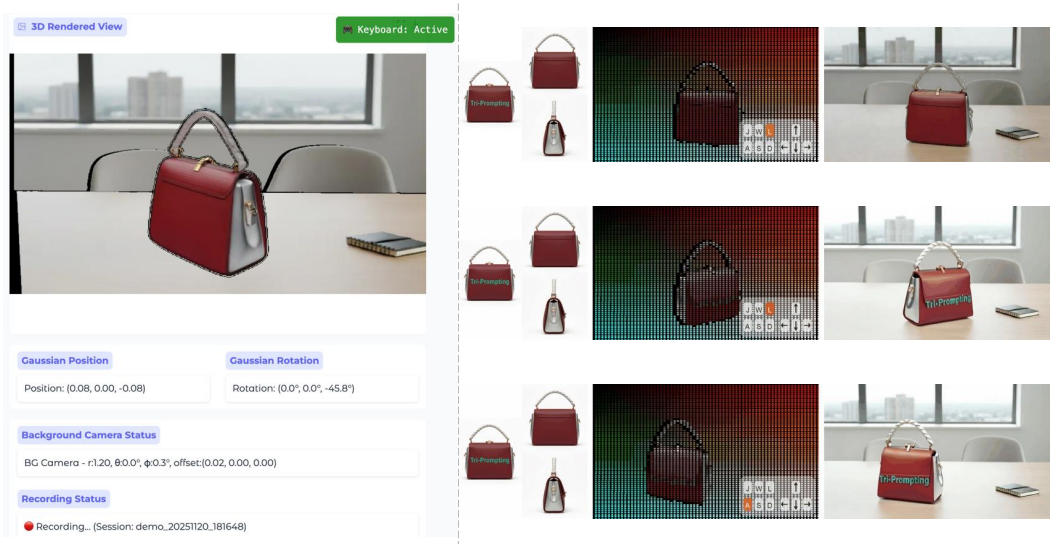


Figure 7: **Offline control interface.** Left: Interface for specifying camera and subject transformations. Right: Examples of subject images, motion control video with keyboard, and generated video.

As mentioned in Sec. 3.4.2 in the main paper and shown in Fig. 7, we design an offline control interface to collect user motion instructions. The user selects a background image and a 3D-aware subject, then controls the background/camera and the 3D subject via the keyboard. We composite the background with the 3D subject's rendering and convert the background into a 70×70 XYZ point and the subject rendering into a low-resolution RGB point, as stated in the main paper. The

interface records the manipulation session; upon completion, the recording is temporally resampled to 49 frames and used as the model's motion control video.

C Implementation Details

Tab. 4 summarizes the training and inference hyperparameters of Tri-Prompting, including the training data, model configuration, and the settings used in the two different training stages. In particular, Stage I fine-tunes the base Phantom 14B model with LoRA for scene and multi-view subject control, while Stage II further trains the ControlNet module for dual-conditioning motion control. We leave joint optimization of LoRA and ControlNet to future work. We also report the inference hyperparameters for the ControlNet scale schedule, namely the decay steps N_{decay} and the minimum control scale $s_{\min}$, which correspond to the inference-time ControlNet scale scheduling strategy described in Sec. 3.4.1. Unless otherwise specified, this schedule is used in both the comparison with Phantom and the application results. We do not apply this strategy in the comparison with DaS, since in that setting both the low-resolution RGB guidance and the XYZ tracking points are directly derived from the original video.

Table 4: **Detailed hyperparameter settings.**

Parameter	Value
Video frames	49
Spatial resolution	832×480
Denoising steps	50
Base model	Phantom 14B
Stage-I Learning rate	$1e-4$
Stage-I batch size	8
Stage-I step	2500
Stage-I LoRA rank	64
Stage-II Learning rate	$1e-4$
Stage-II batch size	32
Stage-II step	2074
Stage-II ControlNet layers	18
ControlNet decay steps N_{decay}	10
Minimum control scale s_{min}	0.005

C.1 Inference Cost

We measure Phantom, Stage I, and Stage II on the same eight A100-80GB GPUs, jointly generating one 49-frame, 832×480 video with 50 denoising steps. Tab. 5 reports results on the same three cases. Timing includes input loading, distributed diffusion, VAE decoding, and video encoding; one-time model loading is excluded. Stage I adds 7.10% wall time over Phantom, and Stage II adds 46.75% over Stage I while enabling foreground/background motion control.

Table 5: **Inference cost on eight A100-80GB GPUs.** Wall time is mean $\pm$ population standard deviation over three cases. Peak memory is allocated/reserved GiB per GPU.

Method	Wall time (s)	Min/video	Total GPU-min	Peak memory
Phantom	147.122 ± 0.109	2.452	19.616	38.578 / 43.252
Ours, Stage I	157.566 ± 0.118	2.626	21.009	39.174 / 42.928
Ours, Stage II	231.235 ± 0.315	3.854	30.831	52.616 / 54.740

D Additional Baseline Comparisons

D.1 OmniVCus Reference Inputs

We evaluate the released OmniVCus [5] checkpoint with text and foreground-mask videos on the same 106 cases and three seeds. The single-reference setting uses its native image-reference input; the three-view setting supplies a fixed montage through the same input. Scene images and camera trajectories are not supplied separately. Tab. 6 reports both settings alongside Tri-Prompting, which uses the first-frame scene image, three subject views, and RGB/XYZ motion cues. The montage improves DINO identity over a single reference; Tri-Prompting has higher scores on all six metrics.

For the 3D consistency comparison in Tab. 2(b), we use 18 rigid cases with six outputs per case (108 outputs per setting). We apply each output's foreground mask before π^3 reconstruction and align the point clouds with a shared-frame Sim(3) transform. We average E_{align} over the 84 outputs successfully reconstructed across eight settings: Phantom, Stage I, Stage II with one, two, or three

references (including both scale schedules for three references), and the two OmniVCus inputs. Outputs are matched by case and fixed within-case output rank.

Table 6: **OmniVCus reference-input comparison.** 318 outputs per setting; higher is better.

Method	Aes.	Smooth	Dynamic	Imaging	$\mathcal{I}_{\text{dino}}$	$\mathcal{I}_{\text{clip}}$
OmniVCus (1-ref)	0.619	0.987	0.619	0.707	0.695	0.845
OmniVCus (3-view montage)	0.618	0.987	0.613	0.703	0.703	0.846
Ours (3-view)	**0.665**	**0.989**	**0.903**	**0.715**	**0.746**	**0.862**

D.2 Motion and Identity Evaluation

We use the single-entity camera-C12 subset of 360Motion [10], held out from 3DTrajMaster's training on the other 11 cameras. Tri-Prompting is evaluated without training on this dataset. We use the released root 6DoF controls for 3DTrajMaster and downsample the released videos to construct Tri-Prompting's low-resolution foreground RGB. On our 106-case set, 3DTrajMaster receives matched global/entity text and 13 root-6DoF tokens derived from recorded position and rotation. Phantom receives the scene image, three subject references, and text, without motion inputs. Our 106 cases comprise 81 dynamic and 25 static backgrounds; subject controls include 59 rotation-and-translation, 38 rotation-only, and 9 translation-only cases.

We extract target and generated subject masks using SAM 3 [6] on public 360Motion and SAM 2 [28] on our evaluation set. Bbox IoU is the intersection area divided by the union area of the tight bounding boxes enclosing the target and generated subject masks. Mask IoU applies the same ratio to the masks themselves. Both metrics are averaged over valid frames. Multi-view identity uses the DINO/CLIP similarity defined in the main text. We evaluate two fixed ControlNet settings across both datasets: constant scale 1.0, or decay from 1.0 to 0.005 over the first 10 of 50 denoising steps.

D.3 Comparison with MotionBooth

MotionBooth [40] uses three subject references and masks, 300-step per-subject personalization, and target 2D boxes. We evaluate its public Zeroscope implementation on the same 106 cases and three seeds; Tri-Prompting and Phantom use the inputs described in Sec. D.2. To match MotionBooth's native setting, all videos are normalized to 24 frames at 576×320 and 10 fps. Tab. 7 reports quality and identity under this protocol. With the same reference count, Tri-Prompting improves $\mathcal{I}_{\text{dino}}$ by 0.175 and $\mathcal{I}_{\text{clip}}$ by 0.091 over MotionBooth.

Table 7: **Comparison with MotionBooth.** 318 outputs per method, evaluated at 24 frames, 576×320, and 10 fps; higher is better.

Method	Aes.	Smooth	Dynamic	Imaging	$\mathcal{I}_{\text{dino}}$	$\mathcal{I}_{\text{clip}}$
MotionBooth-public	0.540	**0.984**	0.349	0.637	0.562	0.753
Phantom	0.625	0.982	0.613	0.670	0.715	0.839
Ours	**0.647**	0.983	**0.918**	**0.673**	**0.737**	**0.844**

E 3D Consistency Across All Subjects

The rigid-subject evaluation in Tab. 2(b) uses shared Sim(3) alignment to static references. For non-rigid subjects, legitimate articulation such as walking also contributes to the alignment error. We extend π^3 reconstruction to all 106 cases and three seeds, including dynamic subjects. Tab. 8 reports reconstruction success over all 318 outputs and E_{align} over the 230 outputs successfully reconstructed by all three methods.

Table 8: **3D reconstruction across all subjects.** Alignment error is averaged over the 230 common successes.

Method	Reconstruction successes	$E_{\text{align}} \downarrow$
Phantom	264/318	0.062
Ours, Stage I	282/318	0.079
Ours, Stage II	**298/318**	**0.057**

Stage II has the highest reconstruction success and lowest alignment error on the shared subset. For dynamic subjects, the multi-view identity metrics in Tab. 2(b) complement geometry by assessing semantic and appearance consistency.

F Limitation

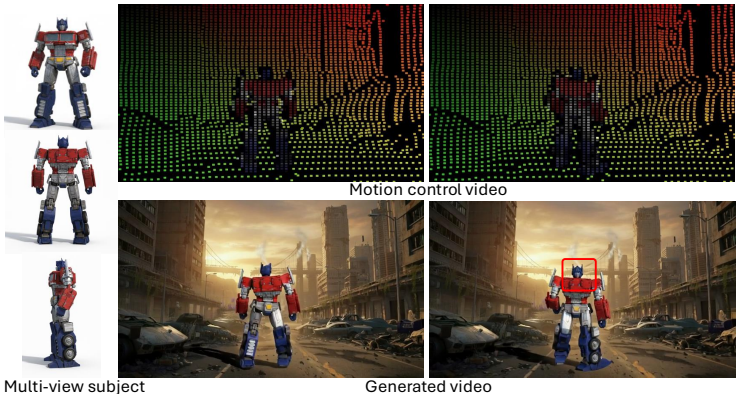


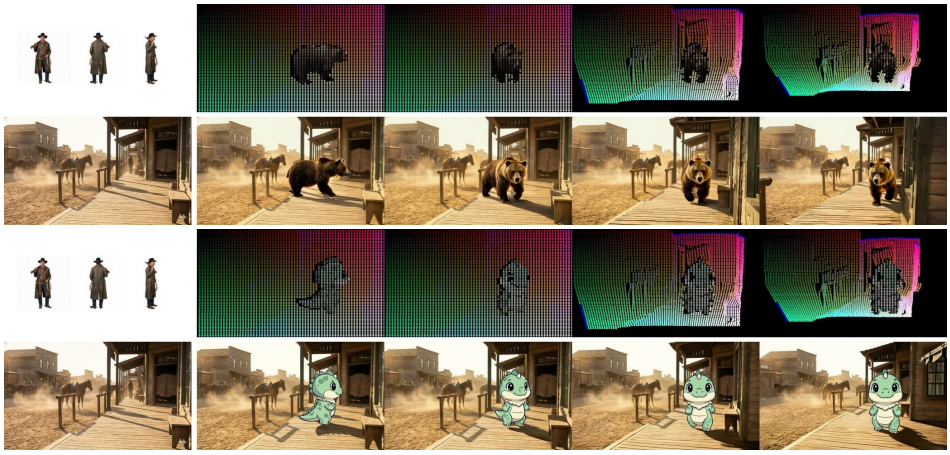
Figure 8: **A failure case of Our Method.**

While Tri-Prompting is generally robust under diverse controls, Fig. 8 illustrates a failure case caused by ambiguity in highly symmetric multi-view references. In this example, the subject's front and back views share very similar appearance, making the identity orientation under motion less distinguishable. As a result, the generated video may exhibit a short-lived inconsistency, such as a transient flip in head orientation during turning. This issue is not persistent across the whole sequence, but it reveals that the current model can struggle when the reference set provides insufficient asymmetric cues. In practice, this failure mode can often be alleviated by using more distinctive reference views or adding stronger motion constraints. In addition, the current method focuses on one primary controllable subject and does not support complex multi-subject generation.

Broader impacts. Tri-Prompting can support controllable video creation and synthetic data generation, but it may also be misused for misleading media, impersonation, non-consensual likeness use, or copyright misuse. Future releases will consider consent requirements, usage restrictions, and provenance or watermarking mechanisms.



a) Across scene

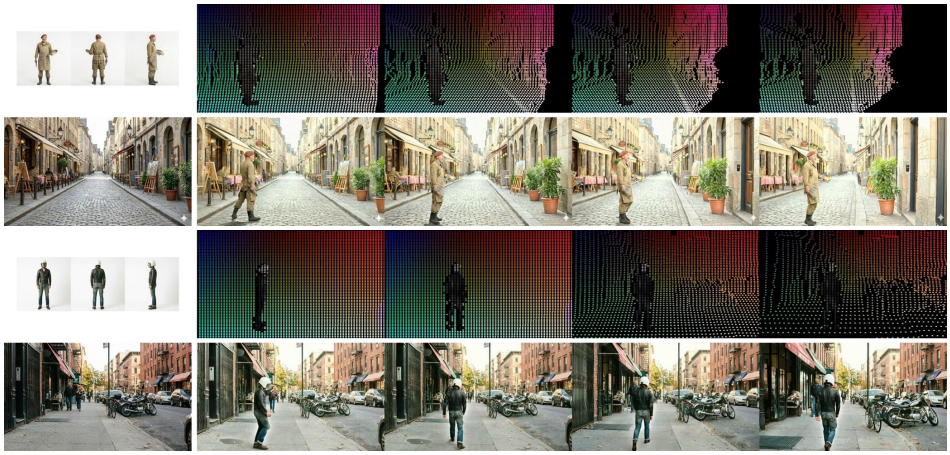


b) Across subject

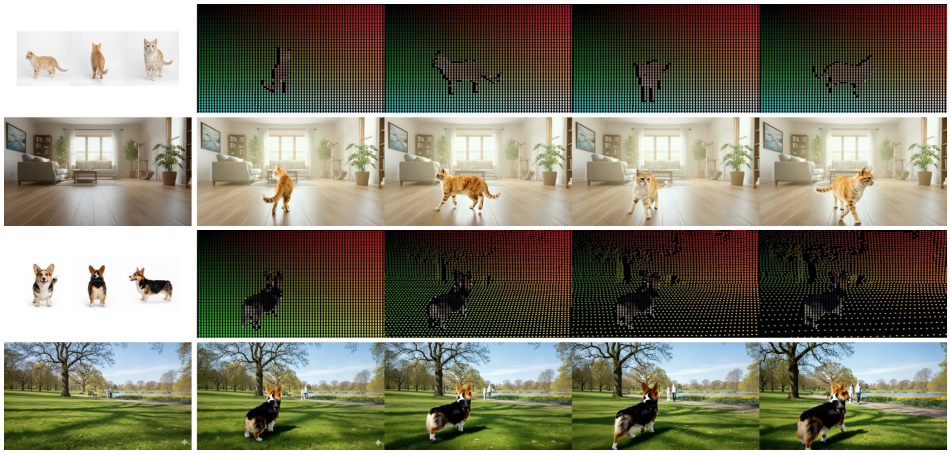


c) Across motion

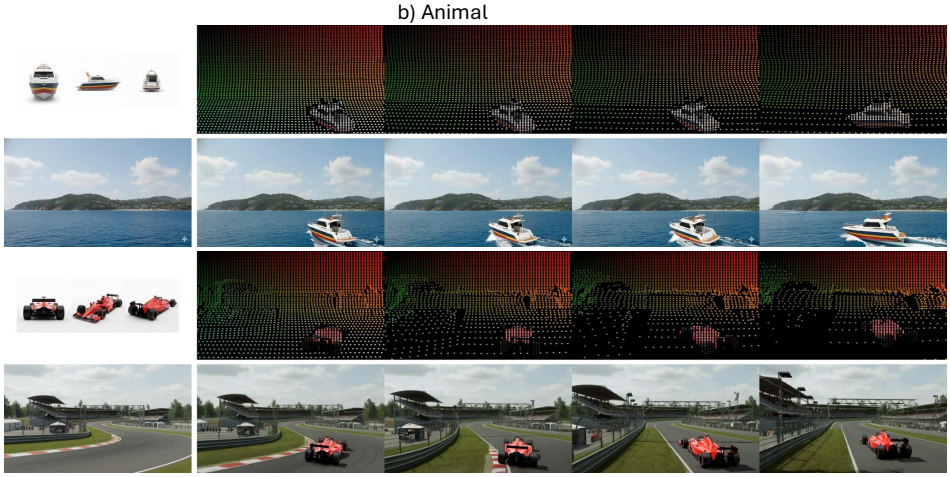
Figure 9: **Subject insertion applications across scene, subject, and motion.** For each example, the left column shows the multi-view subject reference images (top) and the original background image (bottom). The top row visualizes the motion-control video. In the bottom row, the first frame is the scene conditioning image, followed by frames generated by our method.



a) Human



b) Animal

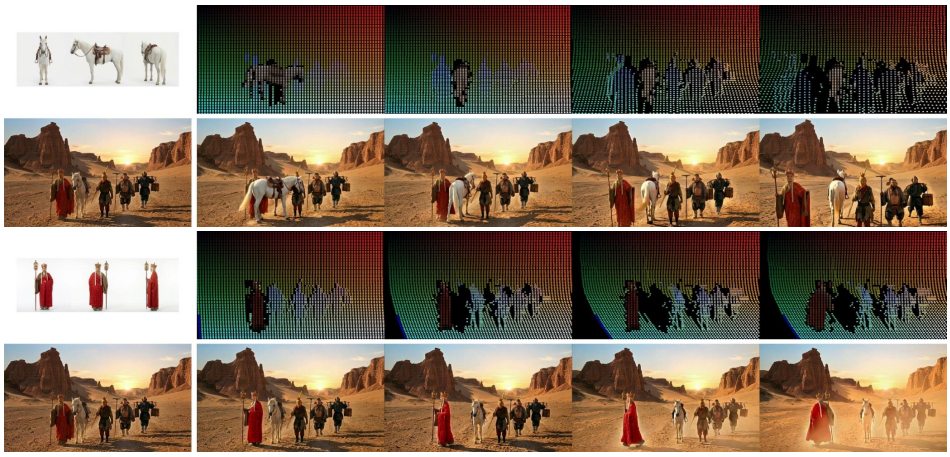


c) Vehicle

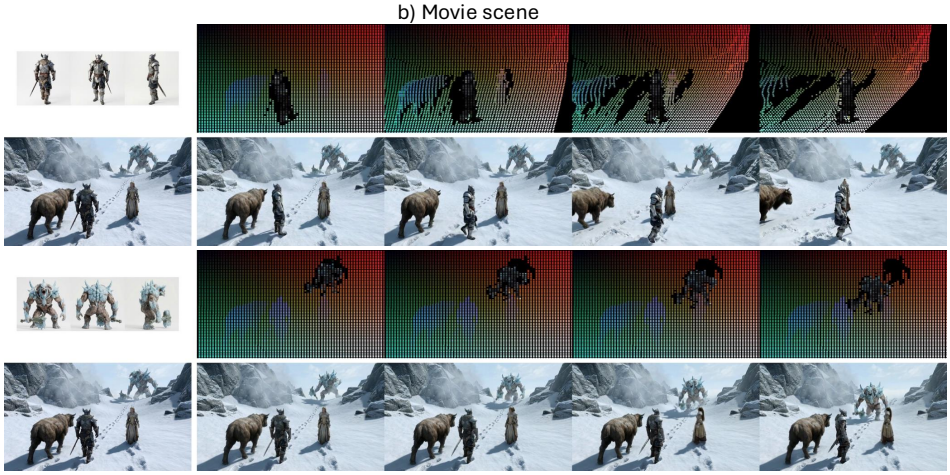
Figure 10: **Subject insertion applications in real-world cases.** Covering both non-rigid subjects (humans/animals) and rigid objects (vehicles). In both categories, our approach achieves high video quality while following the motion instructions.



a) Anime scene



b) Movie scene



c) Game scene

Figure 11: **Subject manipulation applications in anime, movie, and game scenes.** Generalization results across fictional domains, confirming our model's ability to synthesize diverse and varied video output.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's scoped setting: controllable video generation from a first-frame scene image, multi-view subject references, and motion control signals. The main empirical claims are supported by comparisons with DaS and Phantom in Tab. 2 and the ablations in Sec. 4.5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated limitation subsection discussing ambiguity under symmetric multi-view references and the offline, high-latency nature of the current 14B implementation; additional failure discussion appears in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present theoretical results, theorems, or formal proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Sec. 4 describes the training data construction, model backbone, optimization settings, evaluation protocols, metrics, and baseline comparisons; the appendix further reports hyperparameters and inference settings.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper does not currently provide a public release of code, model checkpoints, or processed training tuples. We provide implementation and evaluation details in the paper and appendix to support reproducibility.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Sec. 4 reports the training data sources, tuple construction, model backbone, optimizer, learning rates, batch sizes, training steps, GPU setup, evaluation datasets, and metrics. The appendix further summarizes inference and hyperparameter settings, including resolution, denoising steps, LoRA rank, ControlNet layers, and ControlNet scale scheduling.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We report average results on fixed evaluation sets but do not include formal error bars or significance tests. We compare methods on the same test cases and input conditions where possible.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Sec. 4 reports the GPU type, number of GPUs, training time, batch sizes, and training steps for both stages, and the limitation section reports inference cost for the 14B backbone.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses existing datasets and generative video models for controllable video generation and does not involve human-subject experiments.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The work can support creative video authoring and synthetic data generation, but may also be misused for misleading media, impersonation, non-consensual likeness use, or copyright misuse. The current submission is a research prototype and does not release a public generation service.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A].

Justification: The submission does not release a public model, dataset, or deployed generation service. If artifacts are released in the future, they should be accompanied by appropriate usage restrictions and safeguards against misuse.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the existing datasets, models, and tools used in the paper, use them for non-commercial research and evaluation, follow their reported licenses or terms where explicitly provided, and do not redistribute third-party datasets, checkpoints, game assets, or proprietary editing outputs.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce or release a new standalone dataset, benchmark, codebase, or model artifact as a primary asset. The constructed training tuples are described as part of the experimental procedure.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing experiments or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject research, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as an important, original, or non-standard component of the core method.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.