UNICAMO: UNIFIED CAMERA AND MOTION CONTROL FOR VIDEO DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We present **UniCaMo**, an unified framework for joint control of camera and object motion in image-and-text-to-video diffusion models. Given a reference image, text prompt, sparse 3D object trajectories, and a 3D camera trajectory, UniCaMo constructs motion-consistent input noise that directly encodes the desired motion. Our 3D-grounded noise construction combines reference-derived noise propagation with a spherical Gaussian noise representation for newly revealed regions, providing a simple and effective motion prior without requiring reconstruction of unseen scene geometry. UniCaMo requires no architectural modifications or dedicated control modules and can be adapted using lightweight LoRA fine-tuning with low training data volume. Experiments on standard controllable video generation benchmarks demonstrate state-of-the-art motion controllability and high visual quality with minimal inference overhead. **Qualitative videos are presented on the HTML webpage provided in the submitted supplementary zip files.**

1 INTRODUCTION

Controllable video generation is important for applications ranging from visual content creation to robotics and embodied AI (Wang et al., 2026a; Jang et al., 2025), and has received increasing attention from the research community. Existing approaches generally start from a pretrained image-and-text-to-video diffusion model and finetune it to introduce controllability, broadly following two directions. One line of work adds adaptation or control modules that inject motion guidance into the generation process (Chu et al., 2025; He et al., 2024; Bai et al., 2025a; Li et al., 2025a). Motion constraints are typically injected late in the generation pipeline, when visual content has already begun to emerge, leaving limited flexibility to resolve motion ambiguity in a geometrically coherent manner while maintaining realism and temporal consistency. These methods also augment already large video diffusion models with additional architectural components, increasing model complexity and inference cost. An appealing alternative is to impose motion at the very beginning of generation by exploiting the input noise. Video diffusion models generate videos by progressively denoising an initial sequence of Gaussian noise maps; rather than modifying the model architecture, the desired motion can therefore be encoded directly into these input noise maps (Burgert et al., 2025; Chang et al., 2025).

Existing input-noise-based methods, however, typically rely on 2D optical flow guidance and address object and camera motion separately. These limitations are closely related: the apparent motion of an object in a video is jointly determined by its 3D motion and the changing camera viewpoint, making it inherently ambiguous [1] to specify and control the two independently in 2D image space. Moreover, complex motions—such as one person moving around another (Fig. 1)—are difficult to express without explicit depth and viewpoint information. These observations motivate the central question of this work: *can we jointly control camera and object motions in 3D by manipulating only the input noise, without modifying the architecture of the underlying video diffusion model?*

To answer this question, we propose UniCaMo, a unified framework for joint camera and object motion control through view-and-motion-consistent input noise. UniCaMo lifts the reference image into 3D, where users specify the desired camera motion together with sparse 3D object trajectories. We construct the input noise according to these motions, using reference-derived noise where geometric

[1] The HTML page provides a visualization of motion-noise between existing 2D and our 3D-aware approaches.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

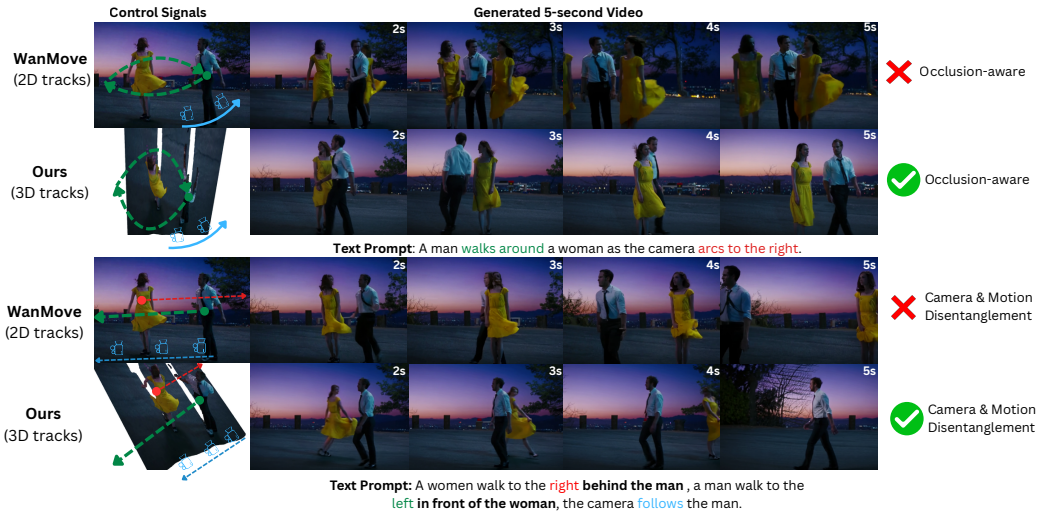


Figure 1: This teaser demonstrates a man walking and dancing around a woman while the camera pans to the right. UniCaMo jointly controls both object and camera motion through a 3D-grounded noise construction, enabling accurate occlusion handling and consistent appearance in unseen regions across frames, an effect that previous work with 2D cues (Chu et al., 2025) cannot achieve.

correspondences are available and a spherical Gaussian noise representation to provide persistent noise for regions newly revealed by camera or object motion. Importantly, this construction need not recover the true geometry of unseen regions; it provides a plausible and geometrically consistent motion prior, while the pretrained video diffusion model supplies the visual and motion priors needed to synthesize realistic content. The resulting sequence of 2D noise maps directly initializes the video diffusion model, enabling joint 3D motion control without additional control modules or architectural modifications. Because the desired 3D motion is already explicitly encoded in the input noise, the model does not need to learn the motion-control mapping entirely from data, allowing effective adaptation with lightweight LoRA fine-tuning and a relatively small amount of computation for training.

Overall, UniCaMo offers three key advantages over existing controllable video generation approaches: **(1)** *unified 3D control with lightweight adaptation*: UniCaMo jointly controls object and camera motion directly in 3D, avoiding the ambiguity of 2D motion specification and separate treatment of the two motions, requiring only lightweight LoRA adaptation; **(2)** *simple and effective motion representation*: its 3D-grounded noise construction provides a plausible and geometrically consistent motion prior without requiring explicit reconstruction of unseen scene geometry; and **(3)** *state-of-the-art performance*: UniCaMo achieves superior controllability and video quality over the evaluated baselines.

2 Related Work

Video diffusion models have made remarkable progress since their early development based on 3D U-Net architectures (Bar-Tal et al., 2024; Blattmann et al., 2023; Chen et al., 2023; Guo et al., 2023). Recent large-scale latent diffusion and transformer-based models (Hong et al., 2022; Kong et al., 2024; Wan et al., 2025; Yang et al., 2024) achieve substantially improved visual fidelity, temporal coherence, and adherence to conditioning signals. As these models become increasingly capable, a natural next step is to provide finer control over the generated motion, particularly the motions of objects and the camera.

Existing motion-controlled video generation methods differ primarily in the motion representation and how it is incorporated into the diffusion model. For object motion, masks, boxes, and keypoints provide coarse spatial guidance, whereas optical flow and point trajectories enable finer-grained control (Wang et al., 2024a; Li et al., 2025a; Geng et al., 2025; Yin et al., 2023; Zhang et al., 2025b). Camera motion is commonly represented by 6DoF poses or trajectories and incorporated through

dedicated conditioning mechanisms (He et al., 2024; Wang et al., 2024b; Bahmani et al., 2025). More recent methods pursue joint object and camera control using trajectory-based conditioning, geometry-aware representations, or richer 3D information (Li et al., 2025a; Geng et al., 2025; Chu et al., 2025; Zheng et al., 2026; Zhang et al., 2026). These approaches generally introduce motion control through additional conditioning signals or modifications within the denoising process.

UniCaMo instead encodes object and camera motion directly into the input noise, connecting our work to structured noise initialization for controllable video generation. Noise-warping methods construct temporally correlated Gaussian noise by propagating noise according to motion fields. Chang et al. (2025) highlight the importance of maintaining appropriate Gaussian noise statistics, while Go-with-the-Flow (Burgert et al., 2025) demonstrates efficient object motion control by warping input noise without an inference-time control module. Related work further shows that training with warped noise can induce useful equivariances and facilitate efficient sampling (Liu & Vahdat, 2025). These approaches, however, primarily operate with 2D motion fields or correspondences. UniCaMo extends this paradigm to joint 3D object and camera control by constructing view-and-motion-consistent noise: noise follows specified object trajectories while remaining consistent across changing viewpoints. A spherical noise representation provides persistent noise for newly revealed regions without explicitly reconstructing unseen scene geometry.

3 UniCaMo – Joint Camera and Object Motion Control

This section describes UniCaMo, our unified framework for joint camera and object motion control in video diffusion through view-and-motion-consistent input noise. UniCaMo requires no architectural modifications to the underlying video diffusion model, enabling lightweight fine-tuning with modest training data while introducing minimal additional overhead at inference time. We first provide an overview of the framework, followed by the details of noise construction and model training.

3.1 Major components and overview

Our overall pipeline is shown in Fig. 2, UniCaMo takes as input a reference image and a text prompt describing the desired scene and action, and outputs a generated video. In between, the framework comprises two stages: (1) an *interactive interface* through which users specify the desired object and camera motions in 3D, and (2) a *motion-guided generation model* that constructs motion-and-view-consistent noise from the reference image and specified motions to initialize an image-to-video diffusion model.

The *interactive interface* first lifts the reference 2D image into a 3D point cloud, allowing the scene to be viewed and interacted with directly in 3D. Users specify object motion using sparse 3D point tracks on objects of interest, together with a desired 3D camera trajectory. This is an important distinction from prior approaches that specify motion using 2D point tracks in the 2D image plane, which are inherently ambiguous with respect to depth and viewpoint and can be insufficient for expressing complex motions, particularly when multiple objects cross or move around one another. By operating directly in 3D, UniCaMo provides a more expressive and geometrically explicit representation for jointly specifying object and camera motion.

The *motion-guided generation model* builds upon a pretrained image-to-video diffusion model (Wan et al., 2025) that takes a sequence of Gaussian noise maps and progressively denoises them into video frames, conditioned on the reference image and text prompt. The input noise and output video therefore have a direct spatiotemporal correspondence: each noise map corresponds to an output frame, and spatial locations in the noise map influence corresponding regions of that frame. Prior work has shown that imposing coherent motion across the input noise maps can induce corresponding motion in the generated video Burgert et al. (2025); Liu & Vahdat (2025). UniCaMo exploits this property by constructing *motion-consistent noise maps* whose cross-frame correspondences reflect the user-specified 3D object and camera motions. Starting from a pretrained video diffusion model, we subsequently fine-tune it to better exploit these motion-consistent noise maps for controllable generation.

The key challenge is to construct consistent noise correspondences across frames given only the reference image and the user's sparse 3D object trajectories and camera poses. Prior methods (Burgert et al., 2025; Liu & Vahdat, 2025) commonly rely on ground-truth video to compute dense optical

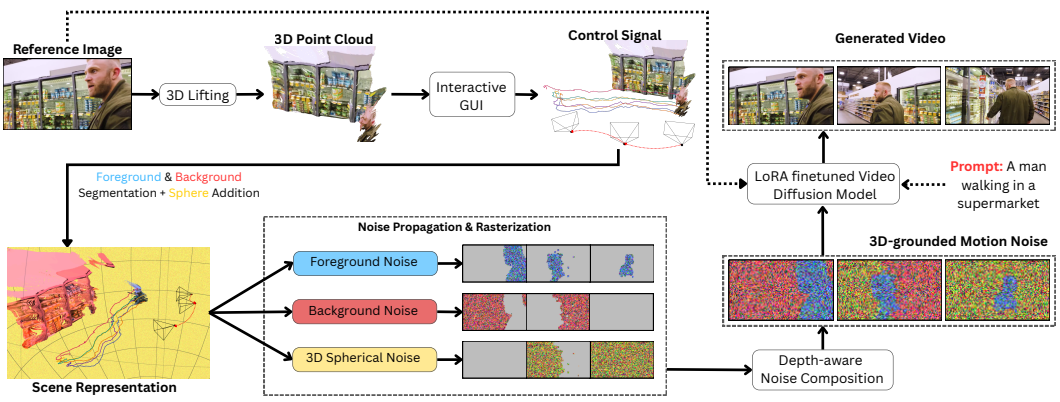


Figure 2: The UniCamo pipeline begins with 3D-lifting the input reference image to a 3D point cloud, within which per-frame control signals are defined via an interactive GUI, producing 3D point tracks and camera trajectories. From these, a 3D-grounded scene representation from three noise components: foreground noise derived from 3D point tracks, background noise from depth cues, and 3D spherical noise for persistent out-of-view anchors, which is then merged by a depth-aware noise composition step. The 3D-grounded noise is consumed by a LoRA finetuned video diffusion model to generate a video exhibiting corresponding object and camera motion.

flow, which creates a systematic training–inference mismatch when such video is unavailable; our approach seeks to eliminate this discrepancy entirely. While noise associated with the reference image can be propagated according to the specified object and camera motions, these motions may reveal regions that are not visible in the reference frame and therefore have no corresponding noise values. UniCaMo addresses this problem by surrounding the reference point cloud with a *3D spherical noise representation*, which provides persistent noise anchors for such unseen regions. Combining the reference-derived and spherical noise allows UniCaMo to construct a complete sequence of 2D noise maps consistent with the desired object and camera motions.

Importantly, the resulting noise sequence represents only *one plausible realization* consistent with the desire motion. Recovering the true 3D geometry of occluded and out-of-view regions from a single reference image is fundamentally ill-posed and, more importantly, unnecessary for our purpose. Instead, the 3D sphere provides a simple and effective construction for assigning persistent noise values to otherwise undefined regions and maintaining their consistency across viewpoints. The sphere therefore does not represent the true scene geometry, nor does it constrain the generated scene to have spherical structure; it merely provides a geometrically consistent motion prior. The final video is determined jointly by this prior, the reference image, the text prompt, and the rich visual and motion statistics learned by the diffusion model from large-scale video data. Thus, rather than explicitly reconstructing unseen scene content, UniCaMo leverages the generative power of the pretrained model to synthesize plausible geometry and appearance while using the structured noise to guide the desired object and camera motions.

3.2 Motion-Consistent Noise Construction

Given a reference image and the desired object and camera motions specified relative to it, our goal is to construct a sequence of 2D Gaussian noise maps whose cross-frame correspondences follow these motions. The motion specification consists of (i) a sparse set of target 3D point tracks $\mathcal{T} = \{\mathbf{x}_i^t \in \mathbb{R}^3 \,|i = \overline{1, N}, t = \overline{1, T}\}$ and (ii) a sequence of target camera poses $\mathcal{P} = \{\mathbf{P}^t | t = \overline{1, T}\}$, where N is the number of point tracks and T is the number of video frames. We first lift the reference image into a 3D point cloud using off-the-shelf 3D reconstruction methods Huang et al. (2025); Li et al. (2025c); Wang et al. (2026b) and associate its points with Gaussian noise sampled for the reference frame. Foreground noise is then propagated according to the specified object trajectories $\mathcal{T}$ and camera poses $\mathcal{P}$, while background noise is propagated according to $\mathcal{P}$ alone. This determines consistent noise values for regions represented by the reference point cloud, but leaves regions that are occluded or outside the reference field of view undefined. To complete these regions, we surround the reference point cloud with a 3D spherical noise representation that provides persistent noise

values across viewpoints. Together, these steps produce a complete sequence of motion-consistent noise maps used to initialize the video diffusion model.

Foreground noise propagation. We obtain a foreground mask of the reference image (Carion et al., 2025) and retain the 3D point tracks in $\mathcal{T}$ associated with the foreground objects of interest. For each frame t, we project each 3D point $\mathbf{x}_i^t$ onto the image plane using the corresponding camera pose $\mathbf{P}^t$ and rescale the projected coordinates to the latent resolution. Specifically, given an image resolution of $W \times H$ and a latent resolution of $W_{\text{lat}} \times H_{\text{lat}}$, we scale the projected coordinates by W_{lat}/W and H_{lat}/H. Let $\mathbf{u}_i^t \in \mathbb{R}^2$ denote the resulting latent-space location of $\mathbf{x}_i^t$. Starting from i.i.d. Gaussian noise sampled for the reference frame, we propagate noise along each trajectory by copying a $(2R-1)\times(2R-1)$ patch centered at $\mathbf{u}_i^1$ in the first frame to its corresponding location $\mathbf{u}_i^t$ in each subsequent frame. We use $R = 2$ by default, corresponding to a 3×3 patch. In this way, foreground noise follows both the specified 3D object motion and the changing camera viewpoint.

Background noise propagation. For the background, we use the 3D point cloud obtained by unprojecting the reference image with its estimated depth and associate each point with an i.i.d. Gaussian noise value in place of its RGB color. The background point cloud remains fixed in world coordinates and is rasterized into each target frame according to the corresponding camera pose $\mathbf{P}^t$. Its projected noise therefore moves solely according to the camera trajectory. Thus, foreground noise follows both object and camera motion, whereas background noise follows camera motion alone.

Spherical noise representation. The foreground and background propagation described above, however, cannot assign noise to regions newly revealed by object or camera motion, as these regions have no corresponding points in the reference point cloud. To provide temporally consistent noise for such regions, we surround the reference point cloud with a 3D sphere centered at the camera center of the reference frame. Let $\mathbf{C}^t \in \mathbb{R}^3$ denote the camera center at frame t, and let $d^0_{\max}$ denote the maximum estimated depth of the reference image. We set the sphere radius to $r = \max\left(\tau \cdot \max_t \left\|\mathbf{C}^t - \mathbf{C}^1\right\|_2, d^0_{\max}\right)$, where $\tau = 1.2$ provides a margin around the camera trajectory. We parameterize the sphere using latitude and longitude and assign its surface a fixed Gaussian noise texture $\mathbf{E} \in \mathbb{R}^{C \times H_{\text{env}} \times W_{\text{env}}}$, where H_{env} and W_{env} are proportional to the video resolution scaled by a factor $k = 4$, and C is the latent channel count of the video diffusion model.

For each frame, we cast a ray from the camera center through each latent location and compute its intersection with the sphere. The intersection point is converted to latitude–longitude coordinates, which are used to sample the corresponding value from the fixed Gaussian noise texture $\mathbf{E}$. Because all frames sample from the same spherical noise texture, rays intersecting the same spherical location receive the same noise value across viewpoints. The sphere therefore provides persistent and view-consistent noise for regions that cannot be associated with the reference point cloud. Details of the ray–sphere intersection are provided in the appendix.

Depth-aware Noise Composition. We finally combine the foreground, background, and spherical noise contributions into a complete noise map for each frame. Multiple foreground and background points may project to the same latent location, particularly under object motion, occlusion, and viewpoint changes. Rather than resolving such collisions through arbitrary ordering (Chu et al., 2025) or averaging (Burgert et al., 2025), we exploit their explicit 3D geometry. For each latent location receiving multiple contributions from the foreground trajectories or background point cloud, we apply a z-buffer rule and retain the contribution with the smallest camera-space depth, i.e., the point closest to the camera. This enforces a geometrically meaningful visibility ordering and prevents noise associated with occluded surfaces from overwriting visible ones. Wherever this propagation provides a valid noise value, we retain it; locations that remain undefined are filled with the corresponding noise sampled from the sphere. This preserves the explicit geometric correspondences available from the reference image while using the spherical representation only where such correspondences are unavailable. The result is a complete sequence of motion noise maps consistent with both the specified 3D object trajectories and camera poses, providing the motion-aware initialization for subsequent video generation.

4 EXPERIMENTS

We evaluate UniCaMo in three settings: object motion control, camera motion control, and joint object and camera motion control. For object motion control, we use MoveBench (Chu et al., 2025),

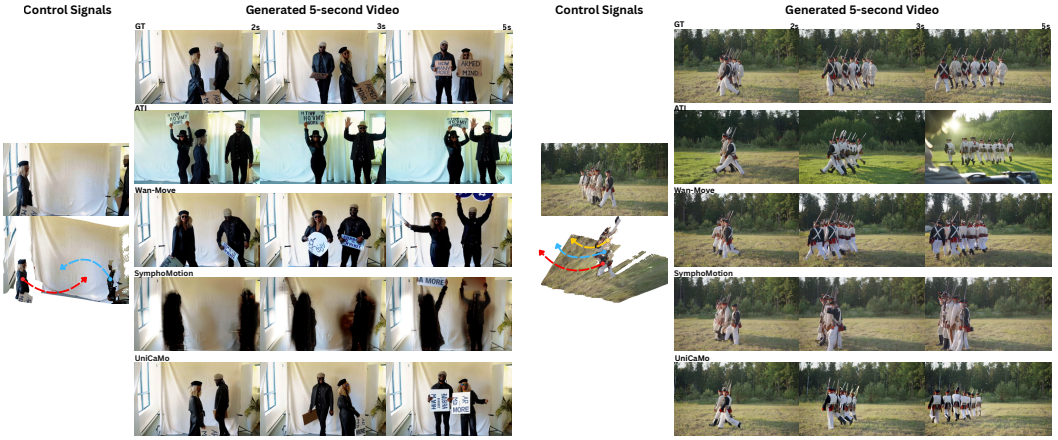


Figure 3: Qualitative results on MoveBench between UniCaMo and recent approaches (Chu et al., 2025; Wang et al., 2025; Zhang et al., 2026).

a recent benchmark comprising approximately 1,000 diverse, high-quality videos. For camera motion control, we randomly sample unseen 300 clips from RealEstate10K (RE10K) (Zhou et al., 2018). Finally, to evaluate joint control, for which no suitable benchmark currently exists, we curate unseen 300 real-world videos from publicly available sources (JovianZM, 2024) spanning diverse object motions and camera trajectories, following the setup in previous work (Chu et al., 2025; Zhang et al., 2026).

Using existing protocol (Chu et al., 2025; Ren et al., 2025; Wang et al., 2024b), we evaluate three complementary aspects of the generated videos: **(i) visual quality**, measured by FID (Heusel et al., 2017), FVD (Unterthiner et al., 2019), PSNR, and SSIM (Wang et al., 2004); **(ii) object-motion fidelity**, measured by the endpoint error (EPE) of the generated trajectories (Chu et al., 2025); and **(iii) camera-motion fidelity**, measured by rotation error (RotErr), translation error (TransErr), and camera motion consistency (CamMC), which quantify the deviation between the generated and reference camera trajectories.

4.1 Comparisons with State-of-the-Art Methods

We compare UniCaMo against a broad range of prior motion-controllable video generation methods, covering object motion control, camera motion control, and joint control. Among these, methods such as Wan-Move (Chu et al., 2025) and SymphoMotion (Zhang et al., 2026) support both object and camera motion control, while methods such as Go-with-the-Flow (Burgert et al., 2025) and ImageConductor (Li et al., 2025b) focus on object motion control, and CamI2V (Zheng et al., 2024) and CameraCtrl (He et al., 2024) focus on camera motion control. We generally evaluate these methods using their publicly released models. To enable more controlled comparisons, however, we additionally construct variants of relevant baselines that match UniCaMo in either the motion specification or the underlying video diffusion model. In particular, *Wan-Move + TAPIP3D tracks* projects the same 3D object tracks used by UniCaMo onto the image plane and provides the resulting 2D tracks to Wan-Move, thereby matching the quality of the input trajectories. *Go-with-the-Flow (Wan2.1-I2V-14B)* uses the same Wan2.1-I2V-14B backbone as UniCaMo and is retrained on the same data with the same training recipe, including the LoRA rank and training schedule, leaving the noise-construction strategy as the primary difference. These variants help disentangle the benefits of UniCaMo's 3D motion specification from those of its motion-and-view-consistent noise construction.

Object motion control. Table 1 reports quantitative results on MoveBench, with results reported separately for single-object and multi-object motion. UniCaMo outperforms prior motion-control methods across all metrics, achieving lower FID and FVD, higher PSNR and SSIM, and lower trajectory error (EPE). The advantage is particularly pronounced on the 192 multi-object cases, where



Figure 4: Qualitative results on RE10K between UniCaMo and other baselines (Zhang et al., 2026; Ren et al., 2025). Our method enable more precise camera control than the baselines.

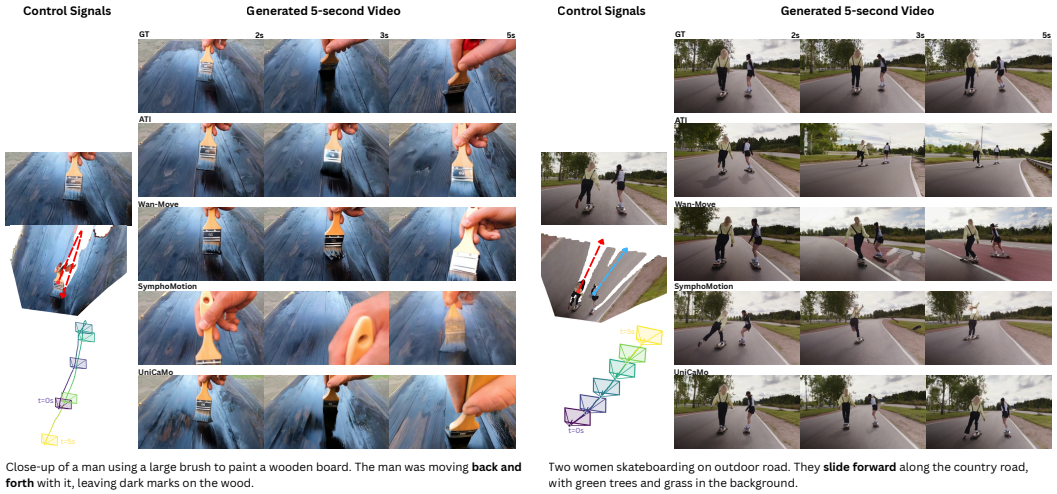


Figure 5: Jointly control over camera and object motions. While other works struggles to generate realistic object dynamics and accurate camera poses, UniCaMo achieves high-quality control.

Table 1: Evaluation results on MoveBench. We report quantitative comparisons on both the single-object and multi-object motion. Best results are shown in **bold**; second-best results are underlined.

Methods	*Test cases with single-object motion*					*Test cases with multi-object motions*				
	FID ↓	FVD ↓	PSNR ↑	SSIM ↑	EPE ↓	FID ↓	FVD ↓	PSNR ↑	SSIM ↑	EPE ↓
ImageConductor (Li et al., 2025b)	34.51	424.11	13.43	0.49	15.66	77.52	764.15	13.92	0.51	9.80
Tora (Zhang et al., 2025b)	22.57	100.24	15.73	0.55	3.30	53.27	350.00	14.50	0.54	3.50
ATI (Wang et al., 2025)	11.27	105.35	15.44	0.58	5.09	30.72	278.91	16.19	0.62	3.17
Go-with-the-Flow (CogX-5B) (Burgert et al., 2025)	12.49	216.90	15.80	0.62	3.04	35.72	399.40	16.93	0.60	3.71
Go-with-the-Flow (Wan2.1-14B)	12.31	96.14	17.60	0.64	2.71	31.06	323.15	17.21	0.63	2.94
Wan-Move (Chu et al., 2025)	12.23	83.50	17.80	0.64	2.60	28.80	226.33	16.70	0.62	2.20
Wan-Move + TAPIP3D tracks (Chu et al., 2025)	12.11	87.10	18.10	0.67	2.54	27.61	294.80	16.50	0.65	2.14
SymphoMotion (Zhang et al., 2026)	**9.92**	89.27	**18.50**	**0.74**	2.33	30.80	351.20	17.57	0.71	2.25
UniCaMo (Ours)	10.36	**78.60**	18.40	0.73	**2.30**	**26.12**	**215.20**	**18.15**	**0.72**	**1.92**

UniCaMo achieves the lowest FVD and EPE, demonstrating stronger adherence to the specified motions in more complex multi-object scenarios. Qualitative comparisons are shown in Fig. 3.

Camera control. Table 2 summarizes the camera-control results. UniCaMo achieves the lowest TransErr (1.89) and CamMC (1.98), while matching the best RotErr (0.26) and achieving FID 5.99 and FVD 39.93. These results support our design of motion-and-view-consistent noise: the spherical noise representation provides persistent noise correspondences across changing viewpoints, including

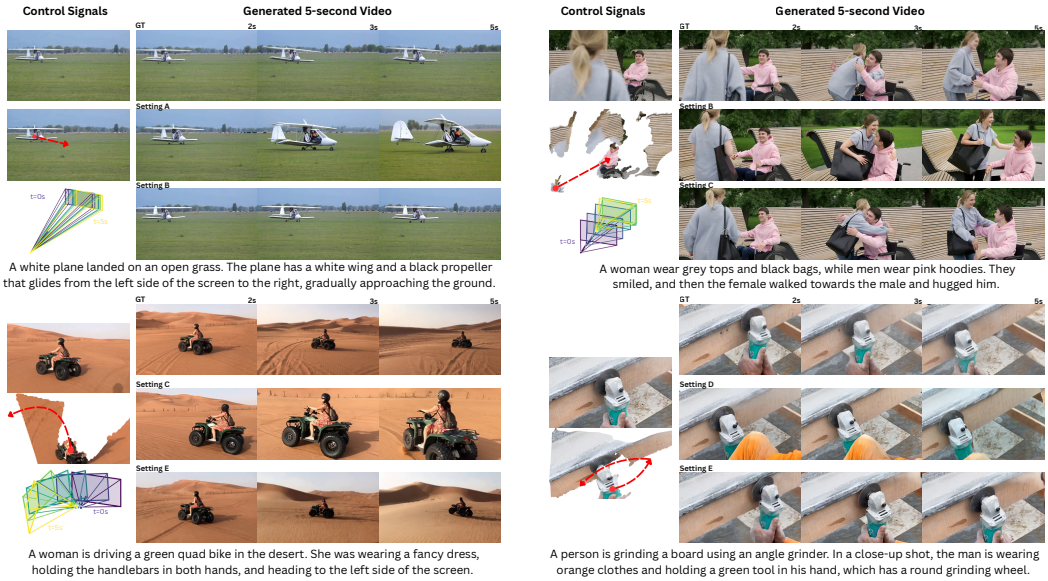


Figure 6: Qualitative results on ablation study of our 3D-aware structured motion noise.

Table 2: Camera-control evaluation on 300 randomly sampled clips from the RealEstate10K test set. We report rotation error (RotErr), translation error (TransErr), camera motion consistency (CamMC), and FID and FVD for visual quality. Best and second-best are **bolded** and underlined.

Methods	FID ↓	FVD ↓	RotErr ↓	TransErr ↓	CamMC ↓
MotionCtrl (Wang et al., 2024b)	23.58	159.52	0.38	7.17	7.23
CamI2V (Zheng et al., 2024)	23.38	167.45	**0.26**	5.15	5.23
CameraCtrl (He et al., 2024)	23.80	165.37	0.27	3.25	3.34
Gen3C (Ren et al., 2025)	6.87	52.39	0.33	2.14	2.28
SymphoMotion (Zhang et al., 2026)	10.33	87.40	1.33	6.31	6.90
UniCaMo (Ours)	**5.99**	**39.93**	**0.26**	**1.89**	**1.98**

Table 3: Evaluation of joint camera and object motion control on 300 clips randomly curated from publicly available sources, each containing simultaneous camera and object motion. Best results are shown in **bold**; second-best results are underlined.

Method	FID ↓	FVD ↓	PSNR ↑	SSIM ↑	EPE ↓	RotErr ↓	TransErr ↓	CamMC ↓
ATI (Wang et al., 2025)	27.83	336.25	14.58	0.54	8.64	18.59	46.41	57.81
Wan-Move (Chu et al., 2025)	23.81	294.49	15.28	0.60	5.03	8.84	44.41	48.49
Wan-Move + TAPIP3D tracks (Chu et al., 2025)	24.17	288.52	15.16	0.58	4.92	8.12	42.11	46.63
SymphoMotion (Zhang et al., 2026)	20.67	219.86	17.95	0.70	3.28	2.34	20.88	21.96
UniCaMo (Ours)	**18.19**	**192.43**	**18.88**	**0.72**	**2.95**	**1.23**	**16.27**	**16.74**

newly revealed regions, enabling the generated video to better follow the specified camera trajectory. Qualitative comparisons are shown in Fig. 4.

Joint camera and object motion control. We evaluate joint control on our 300 curated clips, each containing simultaneous object and camera motion, and report both object- and camera-control metrics in Table 3. We use this curated dataset because MoveBench and RealEstate10K evaluate object and camera motion control separately and therefore do not capture the joint setting for which UniCaMo is designed. UniCaMo achieves the best performance across all eight metrics, with particularly strong gains in this more challenging setting, where the specified object trajectories and camera motion must be satisfied simultaneously. Qualitative comparisons are shown in Fig. 5.

Supplementary videos. Please check our provided HTML page and appendices to see more videos and comparisons.

Table 4: Ablation study of our noise construction method on 300 clips containing simultaneous camera and object motion. Best in **bold**; second-best underlined.

Setting	FG noise	BG noise	Sph. noise	DA comp.	FID ↓	FVD ↓	PSNR ↑	SSIM ↑	EPE ↓	RotErr ↓	TransErr ↓	CamMC ↓
A	✗	✗	✗	✗	21.59	238.41	15.58	0.62	8.23	14.45	43.78	52.86
B	✓	✗	✗	✓	19.38	210.29	17.32	0.66	3.98	7.76	30.45	34.68
C	✓	✓	✗	✓	18.32	**191.51**	18.15	**0.73**	3.16	1.51	18.42	19.11
D	✓	✓	✓	✗	18.51	196.23	18.85	0.71	3.35	1.74	18.85	19.56
E	✓	✓	✓	✓	**18.19**	192.43	**18.88**	0.72	**2.95**	**1.23**	**16.27**	**16.74**

4.2 Ablation Studies

To assess the contribution of each component, we begin from purely random Gaussian noise (Setting A) and introduce our components one at a time, measuring the gain each provides over the preceding setting. Results are reported in Table 4.

Effect of foreground noise. Propagating noise along the foreground 3D tracks (B) show significant improvements over fully independent Gaussian noise (A). It more than halves the trajectory error and substantially reduces every camera metric, while also lowering FID and FVD. Motion written into the initialisation is therefore inherited by the generated video. The top-left of Fig. 6 shows the corresponding qualitative result: random noise still produces a coherent video rather than collapsing outright, but it no longer follows the trajectories specified by the user.

Effect of background noise. Extending propagation from the foreground alone (B) to the background point cloud (C) benefits camera control. This is expected, since the background carries information about scene geometry and layout, and therefore a camera signal that foreground tracks alone cannot supply. The top-right of Fig. 6 illustrates this: Setting B still reproduces the correct object motion, but the camera fails to follow the user's intention.

Effect of spherical noise. Adding the spherical representation for regions absent from the reference point cloud (E) yields a further camera-control gain over C, while FID and FVD remain essentially unchanged, indicating that the gain is specific to camera fidelity rather than to appearance. The bottom-left of Fig. 6 shows where this matters most: when the camera translates over a long distance or rotates through a large angle, extensive novel-view regions are revealed, and the spherical noise supplies most of their content.

Effect of depth-aware noise composition. Resolving collisions with the depth-aware z-buffer rather than compositing in 2D (D) improves performance across the board: EPE drops from 3.35 to 2.95, RotErr from 1.74 to 1.23, TransErr from 18.85 to 16.27 and CamMC from 19.56 to 16.74, with FID and FVD also improving. Ordering contributions by camera depth therefore matters throughout the sequence, not only in explicitly occluded regions. The bottom-right of Fig. 6 shows the failure mode: without depth-aware composition, the motion noise of the grinder is overwritten and washed out by the background, leaving the object static in Setting D.

5 Conclusions

We presented UniCaMo, a unified framework for joint 3D camera and object motion control in image-to-video diffusion models. UniCaMo encodes the desired motion directly into the diffusion initialization by combining noise propagation along 3D object trajectories with a spherical noise representation that maintains persistent noise correspondences across changing viewpoints and newly revealed regions. This view-and-motion-consistent noise enables precise control without additional adaptation modules or architectural modifications to the underlying video diffusion model, requiring only lightweight fine-tuning. Experiments across object motion, camera motion, and joint control demonstrate that UniCaMo achieves strong controllability and visual quality, establishing structured noise initialization as a simple and effective approach to unified 3D motion control in video generation.

AI USE STATEMENT

In this study, generative AI tools were used exclusively to correct grammatical and spelling errors. All AI-assisted edits were reviewed, and every modification proposed by these tools was independently verified by two authors. The authors accept full responsibility for the final content of the manuscript, including any text, claims, or artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

We have provided key algorithms in the main text and the appendices. Full codebase, model checkpoints and datasets will be publicly available upon acceptance.

REFERENCES

Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22875–22889, 2025.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14834–14844, 2025a.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025b.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

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.

Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13–23, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025.

Pascal Chang, Jingwei Tang, Markus Gross, and Vinicius C Azevedo. How i warped your noise: a temporally-correlated noise prior for diffusion models. *arXiv preprint arXiv:2504.03072*, 2025.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

Zhaoxi Chen, Tianqi Liu, Long Zhuo, Jiawei Ren, Zeng Tao, He Zhu, Fangzhou Hong, Liang Pan, and Ziwei Liu. 4dnex: Feed-forward 4d generative modeling made easy. *arXiv preprint arXiv:2508.13154*, 2025.

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

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1–12, 2025.

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

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

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

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025.

Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. Dreamgen: Unlocking generalization in robot learning through video world models. *arXiv preprint arXiv:2505.12705*, 2025.

JovianZM. Pexels. `https://huggingface.co/datasets/jovianzm/Pexels-400k`, 2024. Hugging Face dataset.

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.

Quanhao Li, Zhen Xing, Rui Wang, Hui Zhang, Qi Dai, and Zuxuan Wu. Magicmotion: Controllable video generation with dense-to-sparse trajectory guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12112–12123, 2025a.

Yaowei Li, Xintao Wang, Zhaoyang Zhang, Zhouxia Wang, Ziyang Yuan, Liangbin Xie, Ying Shan, and Yuexian Zou. Image conductor: Precision control for interactive video synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5031–5038, 2025b.

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. MegaSaM: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025c.

Chao Liu and Arash Vahdat. Equivdm: Equivariant video diffusion models with temporally consistent noise. *arXiv e-prints*, pp. arXiv–2504, 2025.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6121–6132, 2025.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Angtian Wang, Haibin Huang, Jacob Zhiyuan Fang, Yiding Yang, and Chongyang Ma. Ati: Any trajectory instruction for controllable video generation. *arXiv preprint arXiv:2505.22944*, 2025.

Boyang Wang, Haoran Zhang, Shujie Zhang, Jinkun Hao, Mingda Jia, Qi Lv, Yucheng Mao, Zhaoyang Lyu, Jia Zeng, Xudong Xu, et al. Robovip: Multi-view video generation with visual identity prompting augments robot manipulation. *arXiv preprint arXiv:2601.05241*, 2026a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026b.

Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. Boximator: Generating rich and controllable motions for video synthesis. *arXiv preprint arXiv:2402.01566*, 2024a.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024b.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023.

Bowei Zhang, Lei Ke, Adam W Harley, and Katerina Fragkiadaki. Tapip3d: Tracking any point in persistent 3d geometry. *arXiv preprint arXiv:2504.14717*, 2025a.

Guiyu Zhang, Yabo Chen, Xunzhi Xiang, Junchao Huang, Zhongyu Wang, and Li Jiang. Symphomotion: Joint control of camera motion and object dynamics for coherent video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11127–11137, 2026.

Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2063–2073, 2025b.

Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024.

Sixiao Zheng, Minghao Yin, Wenbo Hu, Xiaoyu Li, Ying Shan, and Yanwei Fu. Versecrafter: Dynamic realistic video world model with 4d geometric control. *arXiv preprint arXiv:2601.05138*, 2026.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

A USER INTERFACE

To support unified control of both object and camera motion at inference time, we implement an interactive graphical user interface (GUI) that enables users to define 3D object trajectories and 3D camera trajectories through an intuitive two-phase workflow. Fig. 7 shows a visualization of our GUI and results.

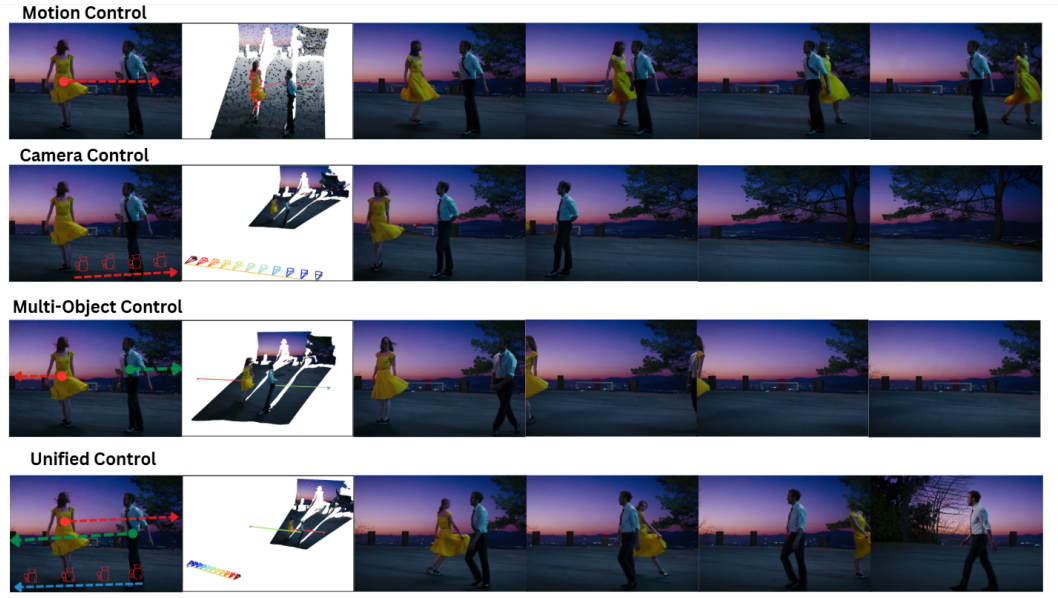


Figure 7: Visualization of our interactive GUI for unified camera and object motion control. Users select objects via polygon masks on the 2D image and edit 3D object trajectories and camera paths using interactive gizmos in the 3D view. Notice how 3D point-tracks allow us to define the exact 3D trajectory of the woman in yellow dress (go behind the man).

Data Processing. Given a single input image, we first run ViPE (Huang et al., 2025) to estimate monocular depth and camera intrinsics. Using the estimated depth map and camera parameters, we reconstruct a dense 3D point cloud of the scene. This point cloud serves as the spatial context for all subsequent editing operations.

2D Object Selection. In the first phase, the GUI displays the input image as a 2D canvas. The user selects the object(s) of interest by drawing polygon masks directly on the image. Multiple objects can be selected as separate groups, each receiving its own mask. From each object mask, we proportionally sample 3D point tracks: for an object whose mask covers n pixels out of the total $H \times W$ image, we allocate $\max(1, \lfloor n/(H \times W) \times N \rfloor)$ of the $N = 1024$ total point tracks. The remaining tracks are sampled uniformly from regions outside all masks to represent background motion. Each sampled pixel is mapped to its corresponding 3D world-space coordinate via the reconstructed point cloud.

3D Object Motion Editing. In the second phase, the GUI transitions to a 3D view showing the reconstructed point cloud alongside the sampled point tracks. The user defines target object trajectories by placing keyframes along a desired 3D path using interactive translation and rotation gizmos (6-DoF transform controls). Intermediate frames are obtained via linear interpolation between keyframes, producing continuous target motion trajectories for all points within each object group.

Camera Trajectory Editing. Camera motion is edited through a separate set of 6-DoF gizmos that control the virtual camera's position and orientation at user-specified keyframes. The camera

trajectory is visualized as a sequence of frustums connected by a spline curve, allowing the user to intuitively design complex camera paths including panning, tilting, dollying, and orbiting. As with object motion, intermediate camera poses are linearly interpolated between keyframes.

Export and Video Generation. Once the user finalizes both object and camera trajectories, the GUI exports a single `.npz` file containing: (i) 3D point track coordinates of shape $(T, N, 3)$, (ii) per-frame visibility flags of shape (T, N), (iii) camera intrinsics of shape $(T, 3, 3)$, and (iv) camera extrinsics (world-to-camera) of shape $(T, 4, 4)$. These exported signals are directly fed into our Guidance-based Noise Warping module (Section 3.2) to construct the 3D-grounded motion-consistent noise, which then initializes the video diffusion model for generation.

B Implementation Details

B.1 Training and inference details

Given a ground-truth training video, we first preprocess it using off-the-shelf models to obtain the signals needed for motion-consistent noise construction. Specifically, we estimate camera intrinsics, extrinsics, and depth using ViPE (Huang et al., 2025), foreground masks using SAM3 (Carion et al., 2025), and sparse 3D point tracks using TAPIP3D (Zhang et al., 2025a). These quantities capture the object and camera motions of the target video and allow us to apply the procedure in Sec. 3.2 to construct a sequence of 2D Gaussian noise maps whose cross-frame correspondences are consistent with those motions. This naturally defines a supervised training pair: the motion-consistent noise sequence serves as the input initialization to the video diffusion model, while the original video serves as the target output. Importantly, the extracted geometric and motion signals are used only to construct the noise initializer and are not provided to the model as additional conditioning inputs.

We curate 400K training clips from DynPose100K++, WSDG-1M, and 4DNeX (Huang et al., 2025; Chen et al., 2025), and precompute and cache all signals required for noise construction. To standardize captions across data sources, we re-caption all clips using Qwen2.5-VL (Bai et al., 2025b), following the Wan-Move prompt-extension protocol (Chu et al., 2025). We fine-tune a pretrained Wan2.1-I2V-14B model (Wan et al., 2025) using flow matching, with the motion-consistent noise sequence as initialization and the reference frame and text prompt as conditioning. We apply LoRA only to the DiT denoiser, while freezing the VAE and all encoders. We use a LoRA rank of 64 with α=1 and train for 10K steps in bfloat16 with gradient checkpointing on 32 NVIDIA H100 GPUs, using a batch size of 1 per GPU, for approximately three days. To improve robustness to different guidance densities, we randomly sample 512–1024 points from the extracted 3D tracks during training. At inference time, we use classifier-free guidance with γ=5.0. UniCaMo introduces no additional model parameters at inference time and incurs only approximately 0.2,s of noise-construction overhead per sample, preserving the runtime efficiency of the base model.

B.2 Implementation Details of Foreground Noise Propagation

Patch copy–paste along trajectories. We initialize the first-frame latent noise as i.i.d. Gaussian,

$$\mathbf{Z}_T^1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{1}$$

For each tracked point i and frame t, let (u_i^t, v_i^t) be its (integer) latent coordinates. To propagate an identifiable noise footprint along the trajectory, we copy a small square patch around the track location in frame 1 to the corresponding location in frame t. Using a patch radius r, we define a patch of $(2r-1) \times (2r-1)$ pixels where the copy follows:

$$\mathbf{Z}_T^t\big[u_i^t + \Delta u,\ v_i^t + \Delta v\big] := \mathbf{Z}_T^1\big[u_i^1 + \Delta u,\ v_i^1 + \Delta v\big]\,. \tag{2}$$

where Δu and Δv are the pixel offsets. For $r = 2$, the patch size is 3×3, and the offset is in $\{-1, 0, 1\}$. By transporting the same local noise pattern across time, motion becomes *visible* to the latents so that visual features in the diffusion process are encouraged to follow the specified trajectory. When multiple trajectories project to the same latent location at time t, naive composition yields ambiguous assignments and can introduce inconsistent motion cues. We then apply the Depth-aware Noise Composition rule described in Section 3.2 to resolve the overlap.

B.3 Implementation Details of Spherical Noise Sampling

Ray construction (pixel → world ray). For each pixel (p_x, p_y) (we use pixel centers), we form a unit ray direction in camera coordinates using intrinsics (f_x, f_y, c_x, c_y):

$$\mathbf{d}_c(p_x, p_y) = \frac{1}{\left\| \left(\frac{p_x - c_x}{f_x}, \frac{p_y - c_y}{f_y}, 1 \right) \right\|_2} \begin{bmatrix} \frac{p_x - c_x}{f_x} \\ \frac{p_y - c_y}{f_y} \\ 1 \end{bmatrix}, \tag{3}$$

and then apply a camera-to-world transformation to determine the ray direction in the world space:

$$\mathbf{d}_w(p_x, p_y) = \mathbf{R}^t \, \mathbf{d}_c(p_x, p_y), \tag{4}$$

where $\mathbf{R}^t$ is the rotation from $\mathbf{T}^t_{c \to w}$. The origin of the ray is at the camera location.

Ray-sphere intersection and spherical UV. Let $\mathbf{S}$ be the sphere center with radius r. Let $\mathbf{o}^t$ be the ray origin and $\mathbf{d}_w$ be the ray direction in the world space. We define the direction from the sphere center $\mathbf{S}$ to the ray origin as $\mathbf{O}^t = \mathbf{o}^t - \mathbf{S}$. We perform ray-sphere intersection by solving a quadratic equation. With $b = \langle \mathbf{d}_w, \mathbf{O}^t \rangle$ and $c = \|\mathbf{O}^t\|_2^2 - r^2$, the (non-negative) discriminant is

$$\Delta = b^2 - c, \tag{5}$$

and we take the exiting intersection (consistent with the camera being inside or near the sphere):

$$\tau = -b + \sqrt{\Delta}. \tag{6}$$

The hit point and outward normal are

$$\mathbf{P} = \mathbf{o}^t + \tau \mathbf{d}_w, \qquad \text{and} \qquad \mathbf{n} = \frac{\mathbf{P} - \mathbf{S}}{r}. \tag{7}$$

We convert the normal to spherical coordinates using longitude–latitude mapping:

$$u = \mathrm{fract}\left(\frac{\mathrm{atan2}(\mathbf{n}_x, \mathbf{n}_z)}{2\pi} + \frac{1}{2} \right), \tag{8}$$

$$v = \frac{\arccos(\mathrm{clip}(\mathbf{n}_y, -1, 1))}{\pi}, \tag{9}$$

where $\mathrm{fract}(\cdot)$ wraps u into $[0, 1)$.

Sampling sphere noise and downsampling to latent. We sample the Gaussian noise map at the computed spherical coordinates using nearest-neighbor lookup, yielding a per-frame noise image. This yields a 3D-consistent noise field while preserving the spatial Gaussianity per frame.

C Additional Ablation Studies

C.1 Effect of various hyperparameters.

Degradation rate. Our motion-consistent noise is spatially Gaussian but introduces cross-frame correlations that differ from the independent Gaussian prior used to train the backbone. We therefore blend it with independent Gaussian noise using a randomly sampled degradation rate $\theta \in [0, 0.5]$. At $\theta = 0$, the initialization preserves the motion signal most strongly, yielding the best control fidelity (EPE 1.88, PSNR 18.35), but also deviates most from the original noise prior, resulting in lower video quality (FID 26.93, FVD 235.3). Increasing θ reduces this mismatch and monotonically improves FID and FVD, at the cost of motion-control fidelity. We adopt $\theta = 0.2$ as a good trade-off between video quality and control fidelity (Table 5 (a) and Fig. 12).

Point tracks. We extract 1024 tracks per video using TAPIP3D (Zhang et al., 2025a) and vary the number used during training to assess robustness to the track budget (Table 5(b) and Fig. 13). Visual quality is relatively insensitive to this choice, with FID and FVD varying by at most 0.4 and 9.4, respectively. Control fidelity is more sensitive: PSNR decreases from 18.15 with 1024 tracks to 16.44 with 512 and 16.12 with 256, while EPE increases from 1.92 to 2.38. We therefore use 1024 tracks by default, although performance degrades gracefully as the number of tracks decreases.

Table 5: Analysis of *(a)* the degradation rate θ for blending the constructed motion-consistent noise with Gaussian noise, *(b)* the number of 3D point tracks, and *(c)* the foreground noise patch dimension, evaluated on MoveBench evaluation data with multi-object motions.

Setting	FID ↓	FVD ↓	PSNR ↑	SSIM ↑	EPE ↓
(a) Degradation rate θ					
$\theta = 0.0$	26.93	235.3	**18.35**	**0.72**	**1.88**
$\theta = 0.2$	26.1	215.2	18.15	**0.72**	1.92
$\theta = 0.5$	**26.03**	**212.78**	17.35	0.67	2.12
(b) Number of 3D point tracks					
$N = 1024$	**26.1**	**215.2**	**18.15**	**0.72**	**1.92**
$N = 512$	26.5	223.3	16.44	0.65	2.18
$N = 256$	26.3	224.6	16.12	0.62	2.38
(c) Patch dimension in Foreground Noise Propagation					
$R = 1.0$	27.6	254.6	15.32	0.54	3.34
$R = 2.0$	**26.1**	**215.2**	**18.15**	**0.72**	**1.92**
$R = 3.0$	26.8	221.19	16.70	0.64	2.28

Table 6: Robustness of UniCaMo to errors in the estimated 3D signals, evaluated on the multi-object split of MoveBench. *(a)* perturbs the per-pixel depth by a controlled percentage; *(b)* substitutes the camera poses and depth of different 3D foundation models. Best result in each block in **bold**.

Setting	FID ↓	FVD ↓	PSNR ↑	SSIM ↑	EPE ↓
(a) Perturbation of the estimated depth					
Depth perturbation = 0% (default)	**26.08**	**215.20**	**18.15**	**0.72**	**1.92**
Depth perturbation = 5%	26.12	215.21	18.11	0.71	1.94
Depth perturbation = 10%	27.14	223.56	17.35	0.68	2.10
Depth perturbation = 30%	31.20	283.32	16.22	0.60	2.57
(b) Source of the estimated camera poses and depth					
MegaSaM (Li et al., 2025c)	26.23	216.40	18.30	0.69	2.23
ViPE (Huang et al., 2025) (default)	26.08	**215.20**	18.15	**0.72**	1.92
VGGT-Ω (Wang et al., 2026b)	**25.93**	215.87	**18.37**	**0.72**	**1.88**

Patch size. We vary the patch dimension $(2R - 1)$ used to propagate noise along the 3D point tracks; results are reported in Table 5(c) and Fig. 14. $R = 2$ performs best across all five metrics. Both a smaller patch ($R = 1$) and a larger one ($R = 3$) degrade performance, suggesting that the patch should be large enough to provide a sufficiently strong motion signal while remaining sufficiently localized to avoid interference between neighboring trajectories. We therefore use $R = 2$ by default.

C.2 Robustness to Errors in 3D Estimation

UniCaMo depends on off-the-shelf modules for depth, camera pose and 3D point tracking, both when constructing training data and when preparing inference guidance, and errors in those estimates propagate into the constructed noise. To make the model tolerant of them, we add a small perturbation to the clean motion-visible noise during both training and inference. We assess that tolerance in two ways on the multi-object split of MoveBench (Table 6).

Perturbed depth. We perturb the per-pixel depth estimated by ViPE (Huang et al., 2025) by a controlled percentage. At 5% and 10% the effect is small — 0.04 FID and 0.02 EPE, and 1.06 FID and 0.18 EPE respectively — which spans the error regime of modern depth estimators. Only at 30% does quality degrade appreciably (5.12 FID, 0.65 EPE). This is expected: a perturbation of that magnitude changes the depth ordering of the scene, and therefore which surfaces occlude which, so generation is guided by a geometrically different scene.

Alternative 3D foundation models. We also replace the estimated camera poses and depth with those of other state-of-the-art 3D foundation models. UniCaMo works with every estimator tested, and its accuracy tracks that of the estimator: FID and EPE both improve monotonically from MegaSaM (Li

et al., 2025c) through ViPE (Huang et al., 2025) to VGGT-Ω (Wang et al., 2026b), which gives our best overall result (25.93 FID, 1.88 EPE). UniCaMo can therefore inherit future advances in 3D estimation without retraining.

D Additional Qualitative Visualizations

In this section, we provide more qualitative results for camera control (see Fig. 8), motion control (see Fig. 9), joint camera and motion control (see Fig. 10). Finally, we show our superior results compared with other baseline on some complex trajectory samples in the Fig. 11. Overall, UniCaMo display precise controllable video generation capabilities even in extreme out-of-domain cases.

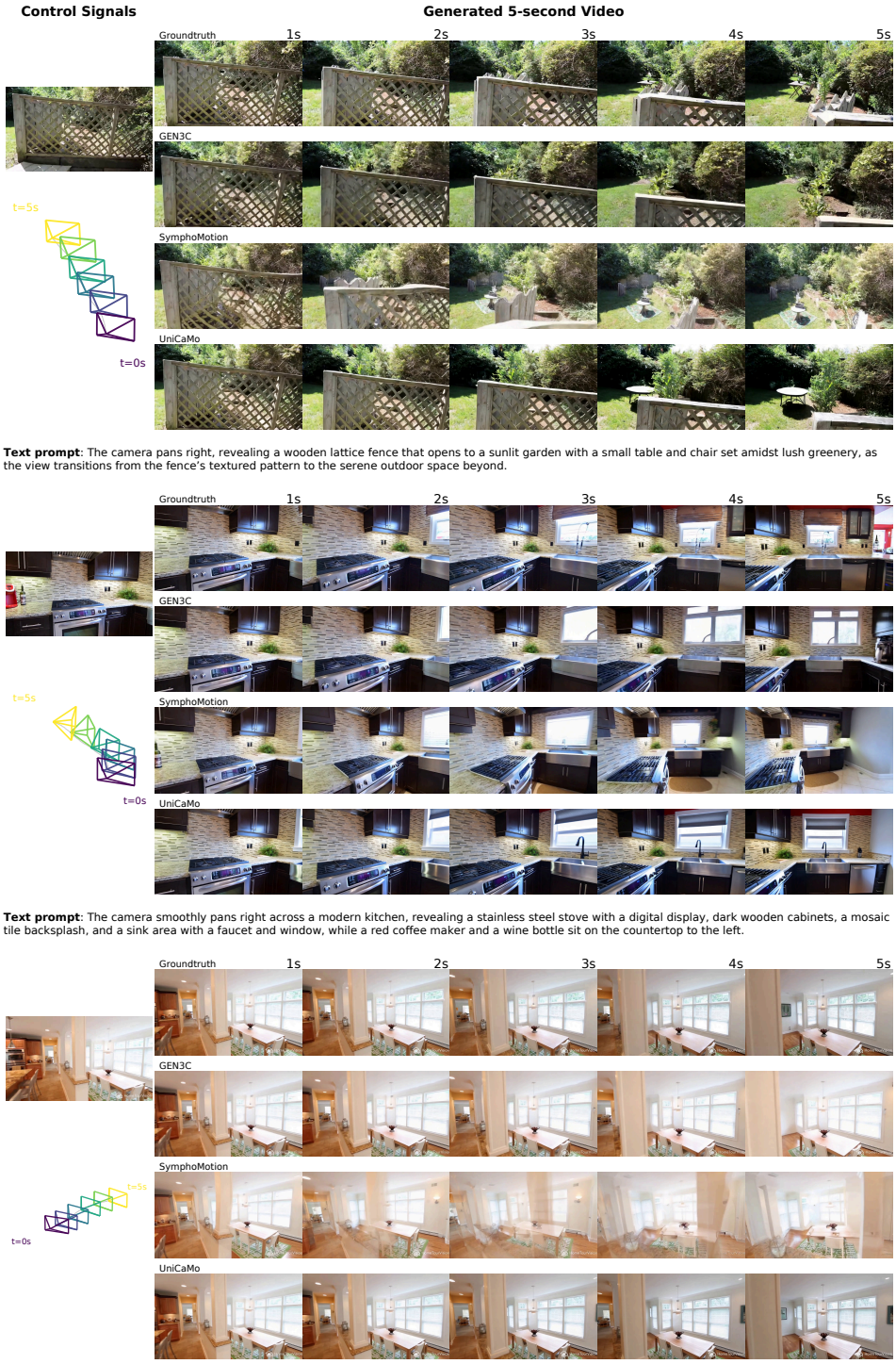
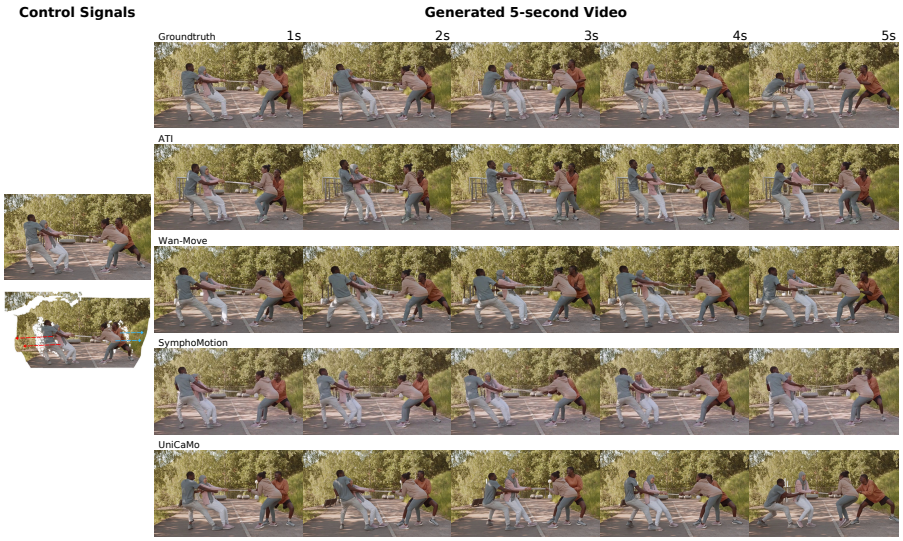


Figure 8: Additional qualitative results for **camera control** on the RealEstate10K dataset.



Text prompt: In the outdoor park, the camera captures a panoramic view of four foreigners playing tug-of-war. Two foreign black men on the left were wearing gray short-sleeved and white pants, and two foreign black women on the right were wearing pink sweatshirts and blue pants.



Text prompt: In the dance studio, the camera flatly captures the medium shot of two foreign white women practicing ballet. They were wearing white training suits, standing in front of the pole, dancing to the rhythm of the music.

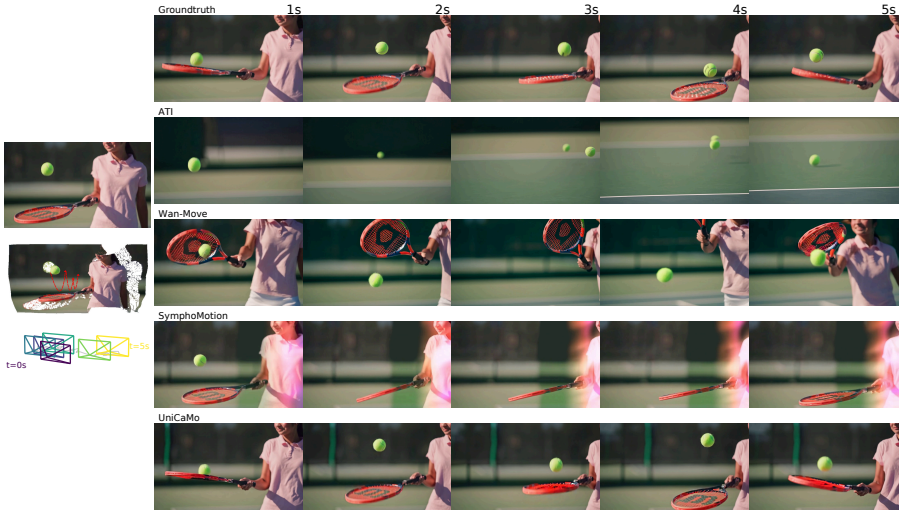


Text prompt: On a snow-white hillside, the camera shot a panoramic view of a group of people skiing. The camera slowly moved backwards, the person in front was wearing red clothes on the black skateboard, and the person in the back was wearing red clothes on the black skateboard.

Figure 9: Additional qualitative results for **motion control** on the MoveBench dataset.



Text prompt: In a spacious and bright hall, a white foreign woman is practicing yoga. In the panoramic flat shot, she was wearing a colorful vest and black pants, sitting on the wooden floor, spreading her legs into a one-line horse pose. Her hair was tied into a tall bun and she wore a necklace as an ornament.



Text prompt: A woman in a pink top is playing tennis on the tennis court. She held an orange tennis racket and was ready to hit the ball. As the video progressed, the tennis ball could be seen falling from the air, and the woman gently touched the tennis ball with her tennis racket to make it bounce.

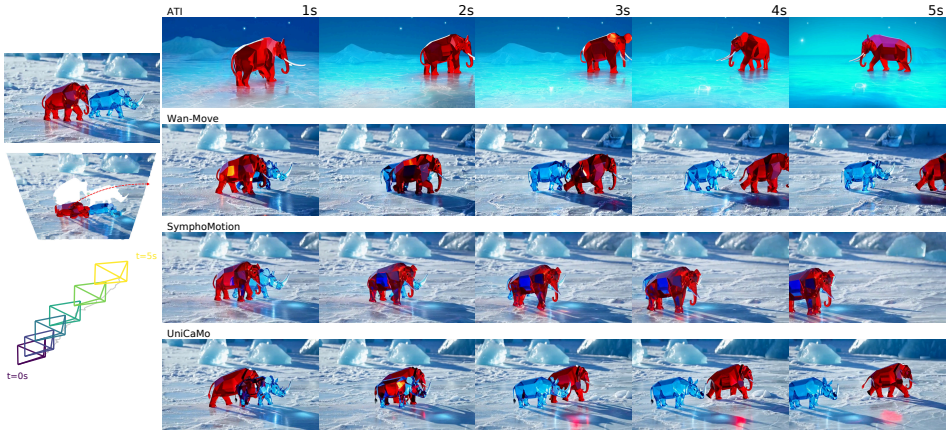


Text prompt: A skier in a yellow ski suit is walking on the snowy mountain. He wore a black hat and pink goggles and held a pink ski with the pattern of the Statue of Liberty. The background is snow-covered mountains and trees.

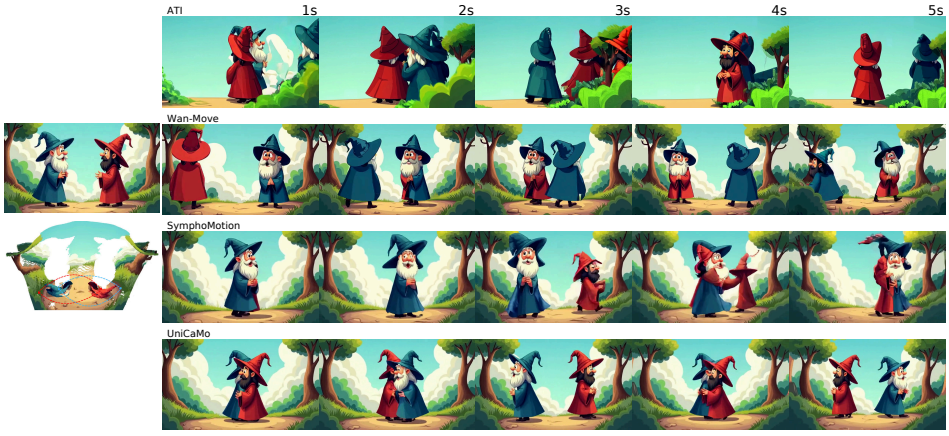
Figure 10: Additional qualitative results for **joint camera and motion control**.



Text prompt: A man walks around a woman as the camera arcs to the right.



Text prompt: A red crystal elephant walk behind the blue crystal rhino, the camera pan right to follow the elephant's movement.



Text prompt: A blue wizard and red wizard walk around each other with the red wizard coming on top first.

Figure 11: Qualitative results for joint control on **copmlex trajectory** cases



Figure 12: Effect of degradation rate on generated samples. A lower degradation rate ($\theta = 0.0$) offers more precise trajectory control but introduces visible artifacts. Conversely, a higher degradation rate ($\theta = 0.5$) yields visually plausible videos but reduces 3D trajectory accuracy. We adopt $\theta = 0.2$ as the default, which achieves the best trade-off between motion fidelity and visual quality.



Figure 13: Qualitative results of our proposed UniCaMo method using different number of 3D point tracks.



Figure 14: Qualitative results of our proposed UniCaMo method using different configuration of patch size R.

E Discussion, Limitations, and Future Work

While our 3D-grounded, motion-consistent noise enables flexible control of object and camera motion, the method has limitations. First, the Foreground Noise Propagation stage depends heavily on accurate 3D point tracks and camera poses, so tracking or calibration errors can degrade controllability. Second, the Spherical Noise Sampling stage uses a coarse 3D sphere proxy, which may fail to capture dynamic background geometry. Future improvements in 3D tracking, depth estimation, and noise representations would further enhance our approach.