Track the Noise, Move the World: 3D-Grounded Motion-Consistent Noise for Controllable Video Generation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Modern image-and-text-to-video diffusion models can synthesize highly realistic videos by iteratively denoising an initial Gaussian noise tensor conditioned on reference image and text inputs. However, existing approaches still lack precise and unified controllability over both object motion and camera motion within a single generation process. We present **UniCaMo**, a unified framework that enables simultaneous control of object trajectories and camera viewpoints by directly constructing the input noise of the diffusion model. Specifically, UniCaMo builds a shared *3D-grounded motion-consistent noise space* across latent video frames. Sparse 3D point tracks are used to warp the Gaussian noise of the reference frame along desired object trajectories, while a virtual spherical noise representation provides globally consistent noise values for newly revealed scene regions under camera motion. By combining local track-guided noise warping with global sphere-based noise sampling, UniCaMo maintains geometric and temporal consistency under both object movement and viewpoint changes. Because UniCaMo modifies only the input noise, it requires no auxiliary adapters, control branches, or architectural changes to the underlying video diffusion model. With lightweight LoRA fine-tuning on large pretrained video diffusion models, including Wan 2.1 (14B), UniCaMo achieves state-of-the-art results in both video quality and motion controllability on standard controllable video generation benchmarks.

1 Introduction

Video is a highly expressive visual medium, and precise control over generated content enables significant practical value. A scene is defined not only by its objects, but by their motion and the camera's movement—e.g., a car on a street looks entirely different when tracked, panned, or viewed from above. Beyond storytelling, controllable video generation is crucial for robotics and embodied AI [41, 25], enabling controllable data synthesis during training and predictive imagination at inference time. Motivated by this, we ask: *can we build a video generation framework that simultaneously and precisely controls object and camera motion without external adapters?*

Controllable video generation has gained increasing attention alongside advances in large-scale video diffusion models [40, 32, 20, 47]. Recent works achieve strong control via point tracks, optical flow, camera trajectories, or auxiliary networks [13, 16, 4, 8, 10, 31]. However, most methods treat object and camera motion separately—either animating objects from a fixed viewpoint or controlling camera motion with unconstrained scene dynamics. When both are specified, this separation often leads to motion ambiguity and entanglement artifacts. Since object and camera motion are inherently coupled, independent control limits expressiveness. In this work, we explicitly tackle joint object-and-camera motion control, enabling a key capability for modern image- and text-to-video frameworks.

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

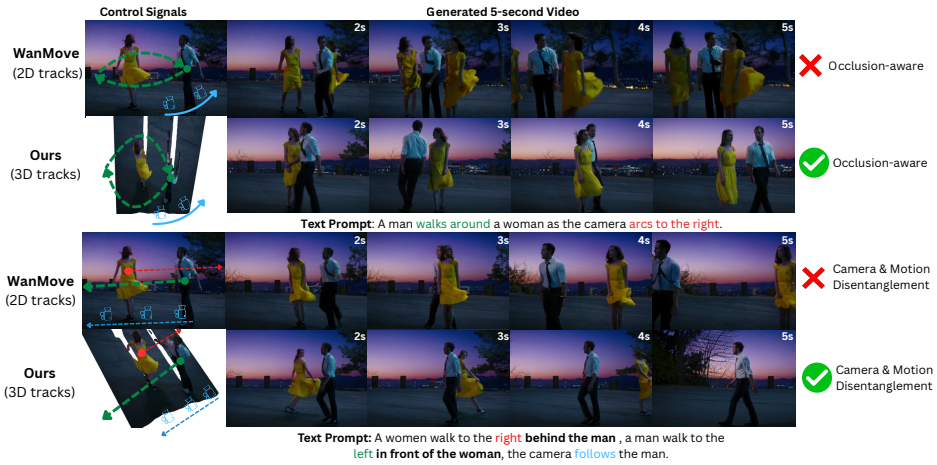

Figure 1: UniCaMo jointly controls object + camera motion with 3D tracks and spherical noise, producing coherent, occlusion-aware videos. WanMove [13] relies on 2D guidance, often causing motion/camera ambiguity and implausible occlusions.

Having established the need for joint motion control, we examine how existing methods specify motion. Prior work [13, 31, 46] typically relies on sparse 2D point tracks, allowing users to drag image-plane points to indicate object or camera motion. While intuitive, this approach has two key limitations. First, 2D trajectories provide an incomplete and ambiguous description of underlying 3D dynamics: the same image-space motion can arise from multiple combinations of object motion and camera movement, and many motions—such as one person moving around another (Fig. 1)—are difficult to express without explicit depth or viewpoint information. Second, motion constraints are usually injected late in the generation pipeline, when visual content has already emerged. This leaves limited flexibility for resolving motion ambiguity in a geometrically coherent way while maintaining realism and temporal consistency. Together, these issues often lead to imprecise, entangled, or geometrically inconsistent motion control.

This observation motivates a different perspective: instead of injecting ambiguous motion guidance near the output, can we impose motion constraints directly at the input? Modern image-and-text-to-video diffusion models generate videos by iteratively denoising an initial Gaussian noise tensor conditioned on a reference image [40]. This suggests a reformulation of controllable video generation as the problem of constructing an input noise tensor that already encodes the desired scene dynamics. Rather than controlling the generation process itself, we control its input. Our approach is thus related to recent noise-warping methods [10, 8], which achieve strong object motion control by warping object-associated noise across frames. However, extending these methods to joint object-and-camera motion is fundamentally challenging. Under a static camera, object motion creates small, localized voids that diffusion models can often infer and inpaint using surrounding context. In contrast, camera motion exposes large unseen regions with no corresponding noise from the reference view. As the viewpoint diverges, these missing regions grow, making noise warping from a single reference-frame fundamentally insufficient for preserving temporal consistency and geometric coherence.

To address these challenges, we propose a motion-consistent noise representation that remains geometrically coherent under both object and camera motion. Instead of defining noise only on the reference image plane, we lift it into a shared spherical 3D representation parameterized by Gaussian noise, providing consistent noise values for all viewing directions. As the camera moves, each frame samples from this shared sphere based on its viewpoint, enabling temporally coherent synthesis of newly revealed regions. Within this space, user-specified point tracks locally warp noise along object trajectories, ensuring consistent object motion under changing viewpoints. This 3D formulation allows users to specify both object and camera motion directly in 3D, avoiding image-space ambiguity. Combining sphere-based noise sampling with track-guided warping, our method—UniCaMo—constructs a unified 3D-grounded motion-consistent noise space that enables precise joint control of object and camera motion while modifying only the input noise. UniCaMo requires no architectural changes or auxiliary adapters, supports lightweight LoRA fine-tuning of

pretrained models such as Wan 2.1 (14B) [40], adds only 0.2s per sample, and achieves state-of-the-art video quality and controllability [13].

In summary, our main contributions are: **(1)** we propose UniCaMo, a novel motion-consistent noise representation that combines 3D point tracks with sphere-projected Gaussian noise to jointly encode object and camera motion in a unified 3D space; **(2)** UniCaMo requires no auxiliary adapters or trainable control modules, introducing only negligible warping overhead (approximately 0.2s per sample) while preserving the runtime efficiency of the base diffusion architecture; and **(3)** extensive experiments on standard controllable video generation benchmarks, including MoveBench [13], demonstrate state-of-the-art controllability and high visual quality, particularly in scenarios involving complex coordinated object and camera motion.

2 Related Work

Diffusion-based video generation has advanced from early 3D U-Nets [6, 7, 11, 15, 19, 18, 34] to large-scale latent diffusion models with stronger spatiotemporal modeling. Transformer-based foundation video models further improve long-range coherence and fidelity via spatiotemporal attention [26, 1, 20, 27, 40, 47, 47, 27]. We build on efficient pretrained models such as Wan [40], and achieve controllable generation through structured latent-noise modeling and lightweight fine-tuning—without architectural changes.

Motion and camera control for video generation. Motion-controlled I2V methods differ by motion cues and how they are injected: masks/boxes/keypoints enables coarse motion cues, while flow or trajectories provide finer guidance [24, 43, 31, 53, 8, 28, 14, 49, 51]. Among these methods, many rely on ControlNet-style adapters with higher cost, whereas Wan-Move [13] edits conditioning features without architecture changes. Camera control in video diffusion is typically achieved by conditioning on 6DoF poses/trajectories, injected via dedicated modules [16, 46, 3]. Related work enforces cross-view consistency (e.g., CVD) or performs camera-only V2V "reshoots," but these methods generally preserve object dynamics and do not support simultaneous object-motion editing [29, 4, 35, 2].

Warped noise for controllable video generation. A promising direction is to control diffusion by using structured initialization noise. Noise-warping methods create temporally correlated noise by warping Gaussian noise with motion fields (e.g., optical flow), injecting temporal structure while preserving per-frame spatial Gaussianity. Warped-noise priors such as HIWYN [10] highlight the importance of this constraint, though large deformations can introduce artifacts and overhead. Go-with-the-Flow [8] makes warped noise efficient and shows motion control can be added by changing only the noise/data pipeline, treating the diffusion model as a black box. Recent analysis further finds that training with warped noise promotes useful equivariances and can enable few-step sampling [33].

Training-free methods. Training-free methods control video diffusion at inference by editing noise/latents, attention, or denoising schedules [21, 48, 37]. Examples include CamTrol and DVS for camera/viewpoint control, and Time-to-Move for motion via region-wise schedules, but such methods often struggle with long-horizon, multi-object, or precise control [21, 48, 37].

Unified control. Existing methods with joint camera and object control often relies on 3D adapters or 3D primitives, increasing runtime cost [43, 31, 14, 36, 13, 30, 52]. Particularly, trajectory-based methods aim for unified control by using background tracks for camera motion and foreground tracks for object motion. Edit-by-Track leverages 3D point tracks to resolve depth/occlusion in joint edits, while other systems use geometry-aware adapters or richer 3D states, typically adding conditioning modules and runtime [30, 52]. In contrast, we target joint camera+object control via structured noise: we add no inference-time adapters [44, 46] and avoid dense flow warping [8] by combining sparse 3D tracks with sphere-projected noise to handle disocclusions while preserving per-frame Gaussianity.

3 UniCamMo: a Method for Joint Camera and Object Motion Control

UniCaMo is built upon an image-and-text-to-video framework such as Wan [40]. It accepts the following input: (i) a source image I and text prompt y, (ii) a sparse set of target 3D point tracks per video frame $\mathcal{T} = \{\mathbf{x}_i^t \in \mathbb{R}^3\}_{i=\overline{1,N},\, t=\overline{1,T}}$, and (iii) target per-frame camera poses $\mathcal{C} = \{\mathbf{P}^t\}_{t=\overline{1,T}}$, with N being the number of point tracks per frame and T the number of video frames. Note that during inference, both $\mathcal{T}$ and $\mathcal{C}$ are defined by the user using a graphical interface.

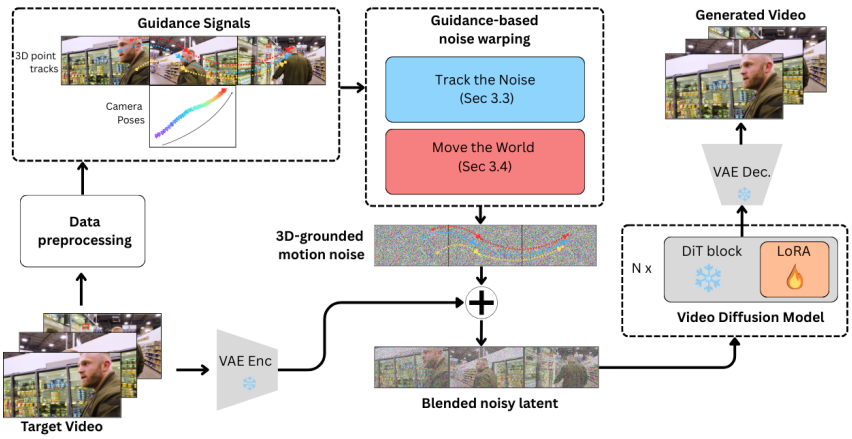

Figure 2: Our framework trains a video diffusion model by warping initialization noise using 3D point tracks and camera poses (via Track-the-Noise and Move-the-World) to create motion-consistent 3D-grounded latents, enabling independent or joint control of object and camera motion at inference.

During training, we construct the input based on original video data as follows. For each video, we extract the first frame and use it as the source image I, and apply video captioning to extract text prompt y. We then apply point tracks, camera pose and depth estimation on the video to obtain point tracks $\mathcal{T}$ and pose $\mathcal{C}$. During inference, we assume that the target 3D point tracks $\mathcal{T}$ and target camera poses $\mathcal{C}$ are specified by the user through a graphical user interface. Our goal is to generate a video with T frames initialized from the content of the source image I with motion dynamics following the target object and camera motion encoded by point tracks $\mathcal{T}$ and camera pose $\mathcal{C}$.

Fig. 2 summarizes our training pipeline. We first perform data pre-processing to obtain depth maps, camera poses, and 3D point tracks of the target video using off-the-shelf 3D models such as ViPE [22] or TAPIP3D [50]. We then apply a noise warping technique to obtain our 3D-grounded motion-consistent noise through two steps: (1) Track-the-Noise (Sec. 3.1), which creates partial motion-consistent noise that captures the motion inferred from 3D point tracks, and (2) Move-the-World (Sec. 3.2), which samples and propagates Gaussian noise on a virtual 3D sphere to produce view-consistent noise along a target camera trajectory. The combined representation results in 3D-grounded motion-consistent noise that can be used to initialize the latents of a video diffusion model [40] to generate the final output video.

3.1 Track the Noise: Noise Initialization using 3D Point Tracks

Motivation. Conventional video diffusion models [40, 47] initialize generation with a single Gaussian noise latent, which effectively behaves as frame-wise uncorrelated noise, lacking any temporal structure. Yet real videos exhibit strong motion correlations, recoverable via dense optical flow [8] or sparse point tracks [30]. Leveraging this as an inductive bias, we partially warp the initial noise using the estimated motion (Fig. 3(a)), producing a motion-consistent noise prior that guides the model toward coherent video dynamics.

Partial Noise Warping. Let $\mathcal{T} = \{\mathbf{x}_i^t \in \mathbb{R}^3\}$ denote the set of world-space 3D point tracks across video frames. Compared to dense optical flow [38], sparse 3D tracks provide explicit depth cues and are robust to occlusion and viewpoint changes. We map the point tracks to latent feature map locations by first projecting them onto the image plane using the target camera poses $\{\mathbf{P}^t\}$, and then rescaling the projected coordinates to the latent feature map resolution. We use the scale ratio W_{lat}/W and H_{lat}/H, where W_{lat}, H_{lat}, W, H are latent and image dimensions, respectively (for Wan 2.1 14B model [40], this scale ratio is 4). Through 3D-to-2D projection, we obtain the 2D pixel trajectories be $\{\mathbf{u}_i^t \in \mathbb{R}^2\}$, corresponding to the 3D point tracks.

By initializing the latent representation of the first frame as i.i.d. Gaussian noise, we now propagate the noise to subsequent frames following the pixel trajectories $\{\mathbf{u}_i^t\}$ constructed from 3D point tracks. We simply transfer (i.e., copy and write) noise within a square patch with dimension $2R-1$ centered at each pixel in the trajectories from the first frame to the target frames. By default, we set $R = 2$,

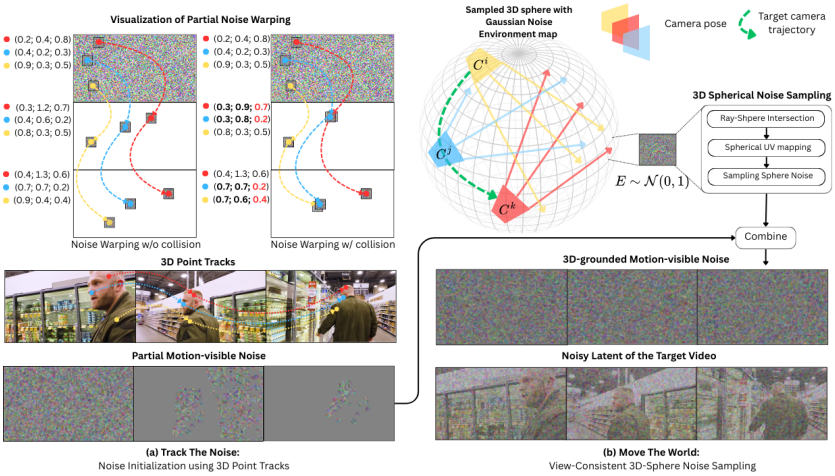

Figure 3: Noise representation in our two key steps, Track the Noise (a) and Move the World (b). (a) We track 3D points (see example triplets of X, Y, Z values) and resolve potential collisions in their trajectories using a depth test, favoring contributions from points closer to the camera. This step results in initial partial motion-visible noise. (b) We create a virtual 3D sphere with Gaussian-textured noise and sample view-consistent noise according to the target camera poses. By combining the outputs of (a) and (b), we obtain 3D-grounded, motion-consistent noise that replaces the randomly sampled noise used in standard video diffusion models.

resulting in patch of size 3x3. As the 3D point tracks encode motion in the video, this results in a video latent representation that exhibits motion-consistent noise, forming a strong guidance for generating motion dynamics during video synthesis.

Depth-aware Collision Tracking. When multiple point trajectories overlap and project to the same latent pixel, naive composition leads to ambiguous assignments and may introduce inconsistent motion cues. We therefore resolve collisions using a depth-aware z-buffer rule induced by the target tracks $\mathcal{T}$ and camera poses $\mathcal{C}$. Among all candidates mapped to a latent pixel, we select the contribution with the smallest camera depth (i.e., the point closest to the camera) and discard the rest. This choice enforces a physically plausible occlusion ordering and leverages the explicit depth information provided by 3D tracks to disambiguate visibility during noise propagation.

In contrast, prior motion-control pipelines often adopt simple but less effective heuristics for handling overlaps, such as selecting one contributor without geometric reasoning (e.g., by arbitrary ordering) [13], or aggregating multiple contributors through averaging (sometimes followed by re-normalization to restore Gaussian statistics) [8]. By explicitly enforcing depth-consistent compositing, our construction yields geometrically coherent structured noise signals, improving the stability of the noise representation under occlusions and viewpoint changes.

3.2 Move-the-World: View-Consistent 3D-Sphere Noise Sampling

Motivation. As the Track-the-Noise step transports first-frame noise along point-track trajectories, some regions remain unconstrained—especially areas that become newly visible as camera motion reveals previously unseen parts of the scene. Filling these gaps with random Gaussian noise provides no motion cues, while copying first-frame noise introduces incorrect parallax. To address this, we introduce Move-the-World, which represents the scene as a 3D bounding sphere coated with Gaussian noise. The camera moves inside this sphere, and projecting the sphere's noise into each view produces view-consistent latents that preserve spatial Gaussianity (Fig. 3(b)).

3D Sphere Parameterization. Constructing a bounded 3D sphere requires defenition on the center position and radius. In particular, we set the sphere center to the camera center of the first frame and choose a radius r such that the sphere encloses the entire camera trajectory across frames while satisfying a conservative depth bound $d_{\max}$. The radius r is define as follows: $r = \max\left(\tau \cdot \max_t \left\|\mathbf{C}^t - \mathbf{C}^1\right\|_2, d^0_{\max}\right)$ where $\mathbf{C}^t \in \mathbb{R}^3$ denote the camera center at frame t,

$d^0_{\max}$ is the furthest depth value of the input image and $\tau = 1.2$. We represent the sphere surface by a latitude-longtitude parameterization, which can be captured by a randomly sampled Gaussian noisy texture map $\mathbf{E} \in \mathbb{R}^{C \times H_{\text{env}} \times W_{\text{env}}}$, where H_{env} and W_{env} are proportional to the video resolution scaled by a factor k ($k = 4$), and C matches the latent channel count of the video diffusion model [40, 47].

Ray-sphere Intersection & Noise Sampling. To obtain 3D-consistent noisy latents for each camera view, we cast rays from the camera through its view frustum and compute their intersections with a bounding sphere using a closed-form ray–sphere intersection formula. For every ray that hits the sphere surface, we determine the latitude and longitude of the intersection point. These spherical coordinates are then used to perform nearest-neighbor grid sampling on the previously defined noisy texture map E. This procedure produces noise latents that remain spatially Gaussian within each frame while staying consistent across different viewpoints. Additional details on the ray–sphere intersection computation can be found in the appendix section.

Noise Composition. To combine the motion-consistent noises produced independently by Track-the-Noise and Move-the-World, we apply a composition step that aggregates their outputs into a single final noise latent. We begin by initializing the final latent with the partial motion-consistent noise obtained from the Track-the-Noise stage. Then, for regions where this step provides no coverage, we fill in the missing areas by copying the corresponding values from the Move-the-World output. This strategy yields a complete latent that is both 3D-grounded and motion-consistent, making it well-suited for the subsequent video generation with the diffusion model.

3.3 Training and Inference

Training data. We curate 400K training clips from DynPose100K++, WSDG-1M, and 4DNeX [22, 12]. For each clip, we precompute and cache offline the signals needed for noise construction: camera intrinsics/extrinsics and depth from ViPE [22], and sparse 3D point tracks from TAPIP3D [50]. These signals are used only to build the motion-consistent noise initializer (not as additional conditioning inputs). To standardize captions across sources, we re-caption all clips with Qwen2.5-VL [5], following the Wan-Move prompt-extension protocol to improve prompt consistency [13].

Training details. We fine-tune a pretrained Wan I2V backbone [40] using flow matching to predict the velocity field that transports noise samples to the data distribution. We start from Wan2.1-I2V-A14B weights and apply LoRA only to the DiT denoiser, freezing the VAE and all encoders. We use LoRA rank 64 with α=1, and train for 10K steps in bfloat16 with gradient checkpointing on 32 NVIDIA H100s (batch size 1/GPU) for 3 days. To improve robustness to varying guidance density, we randomly sample 512–1024 points from the extracted 3D tracks during training.

Inference. Given a single input image, we estimate monocular depth to obtain a 3D point cloud. The user selects target objects via a mask (manual or SAM [9]), then specifies 3D point tracks for these objects over time and optionally a camera trajectory (e.g., via a simple 3D GUI). These signals are passed to our Guidance-based Noise Warping module to construct 3D-grounded, motion-consistent noise. We generate videos with classifier-free guidance ($\gamma = 5.0$). UniCaMo adds no extra parameters and incurs only 0.2s warping overhead per sample, preserving the base model's runtime. Additional qualitative results and demos are provided in the supplement/project page.

4 Experiments

4.1 Experimental Setup

Baselines. We evaluate UniCaMo on controllable I2V generation, focusing on joint camera–object motion. We compare against adapter-based controllers (MagicMotion, Tora, ImageConductor, LeviTor), feature-warping Wan-Move, and flow-based noise-warping Go-with-the-Flow [31, 51, 44, 42, 13, 8]. All methods use the same inputs and we report averages over full benchmark splits.
Benchmarks and Metrics. We report results on the standard controllable video benchmark, namely MoveBench [13]. We evaluate along two axes: **(i) visual quality** using FID [17], FVD [39] and reconstruction-style metrics PSNR, SSIM [45], and **(ii) control fidelity** using motion accuracy metrics such as end-point error (EPE) for trajectories. MoveBench [13] provides only 2D tracks and first-frame masks, so we recover the required 3D tracks and camera poses using off-the-shelf TAPIP3D and ViPE [50, 22]. Using the masks, we sample M object points (proportional to mask size) and $N - M$ background points to match our training track count. To provide an apple-to-apple

Figure 4: Qualitative results on MoveBench between UniCaMo and recent approaches [13, 31, 51]. Our method consistently outperforms other baselines and manages to follow the target motion and camera. Visual artifacts are shown in red boxes.

comparison under matched track sources, we additionally evaluate a *Wan-Move + TAPIP3D tracks* variant, in which we project the same 3D tracks used by UniCaMo onto the 2D image plane and feed them as input to Wan-Move.

4.2 Qualitative and Quantitative Evaluation

Figure 4 shows representative examples in MoveBench under joint camera and object motion control. Overall, UniCaMo generates temporally coherent videos that better preserve scene structure while following the specified object trajectories and camera motion. Compared to another baselines, our results exhibit fewer visuals artifacts and manages to follow the conditioning camera and object motions. Go-with-the-flow [8] produces visible artifacts in small detail like human's heah or feet, Wan-Move [13] while exhibits less artifacts in small detail, failed to generate motion that follow user's intention. Please see the included videos in the supplementary material for more scene details.

Figure 5 demonstrates the generalization of our method on unseen in-the-wild samples with complex geometry and trajectory. In the first example, we expect the white van to move leftward out of the video frame while the camera moves to the right. In this case, Wan-Move [13] failed to generate plausible results. Our 3D-grounded motion-consistent noise representation represents such motion dynamics based on user-provided point tracks and camera trajectories, resulting in high-quality motion content. In the second example, similarly, Wan-Move can pan the camera to the right but does not generate the cars in the correct moving direction. Contrastively, our method can generate realistic flows of the cars in the crowded street.

Single-object motion control. Table 1 (left block) summarizes quantitative results on the single-object split of MoveBench. UniCaMo achieves the best overall performance across both video quality and control fidelity metrics. Our model yields substantially improved quality and control compared to prior motion-control methods, with lower FID/FVD and higher PSNR/SSIM while reducing trajectory error (EPE). In particular, on MoveBench we obtain FID 10.36 and FVD 78.6, improving over Wan-Move (FID 12.2, FVD 83.5) and trajectory-based baselines in the same evaluation protocol. We also observe consistent gains in SSIM and EPE (e.g., SSIM 0.73 and EPE 2.3 for our model versus SSIM 0.64 and EPE 2.6 for Wan-Move). Notably, Go-with-the-Flow [8] did not produce

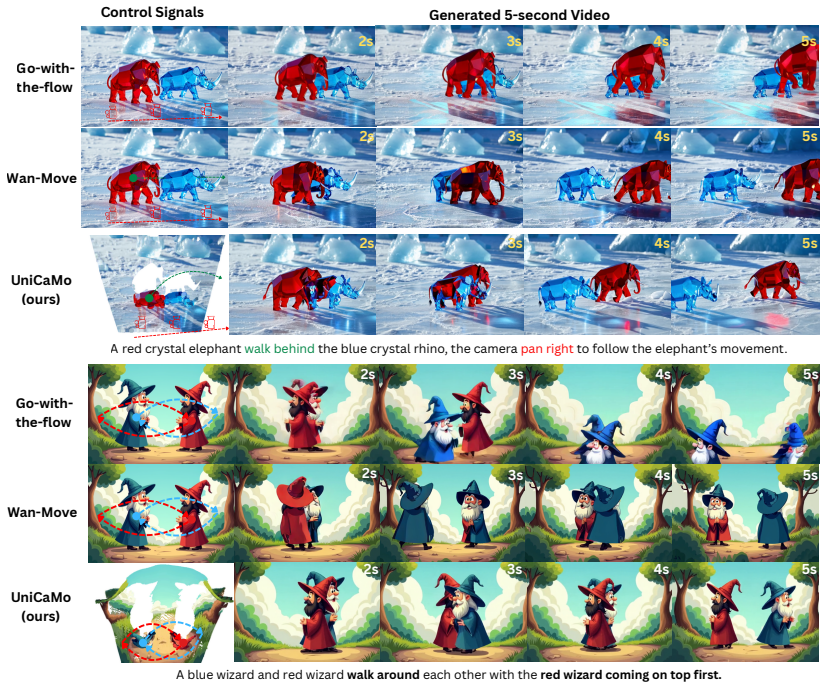

Figure 5: Qualitative results on complex in-the-wild samples between our proposed UniCaMo and Wan-Move [13], Go-with-the-flow [8]

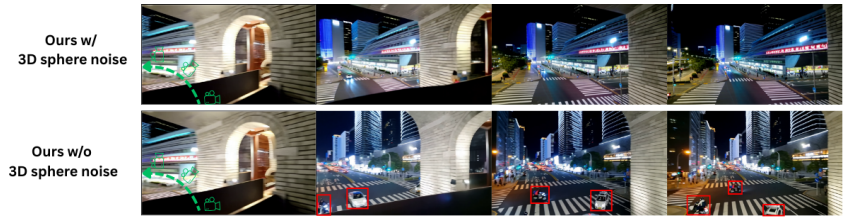

Figure 6: Ablation results on the 3D sphere noise. Without 3D-consistent sphere noise sampling, the model tends to hallucinate random objects (red boxes) with implausible motion trajectories.

Table 1: Evaluation results on MoveBench. We report quantitative comparisons on both the single-object and multi-object motion splits. "–" indicates the baseline does not provide results for that split. UniCaMo achieves best overall performance across. Best in **bold**; second-best underlined.

Methods	*MoveBench (Single-object motion)*					*MoveBench (Multi-object motion)*				
	FID ↓	FVD ↓	PSNR ↑	SSIM ↑	EPE ↓	FID ↓	FVD ↓	PSNR ↑	SSIM ↑	EPE ↓
ImageConductor [44]	34.51	424.1	13.4	0.49	15.66	77.5	764.5	13.9	0.51	9.80
LeviTor [42]	18.12	98.8	15.6	0.54	3.40	–	–	–	–	–
MagicMotion [31]	17.53	96.7	14.9	0.56	3.20	–	–	–	–	–
Go-with-the-Flow [8]	12.49	216.9	15.8	0.62	3.04	35.7	399.4	16.9	0.60	3.71
Tora [51]	22.57	100.4	15.7	0.55	3.30	53.2	350.0	14.5	0.54	3.50
Wan-Move [13]	12.23	83.5	17.8	0.64	2.60	28.8	226.3	16.7	0.62	2.20
Wan-Move + TAPIP3D tracks [13, 50]	12.11	87.1	18.1	0.67	2.54	27.6	294.8	16.5	0.65	2.14
UniCaMo (Ours)	**10.36**	**78.6**	**18.4**	**0.73**	**2.30**	**26.1**	**215.2**	**18.15**	**0.72**	**1.92**

meaningful videos under MoveBench's joint-control protocol despite repeated attempts, indicating that flow-based warped-noise priors do not transfer to this evaluation.

Multi-object motion control. As MoveBench includes 192 cases with annotated multi-object motion, we further evaluate UniCaMo against baselines [44, 51, 8, 13] on this challenging setting, as presented in Table 1 (right block). Our method achieves the lowest FVD and reduced EPE compared to other methods, highlighting its precise adherence to motion constraints in more complex scenarios.

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

Table 2: Ablation studies on UniCaMo on the MoveBench [13] single-object motion evaluation set. We report VBench-I2V [23] scores together with PSNR, SSIM, and End Point Error (EPE). *(a)* compares our method with and without the sphere-projected noise of Move-the-World; *(b)* sweeps the degradation rate θ used to blend constructed motion-noise with pure Gaussian noise.

Setting	PSNR↑	SSIM↑	EPE↓	Subject Consistency↑	Background Consistency↑	Aesthetic Quality↑	Imaging Quality↑	Overall Consistency↑	Temporal Flickering↑	Motion Smoothness↑
(a) Effect of sphere-projected noise (Move-the-World)										
Ours w/ sphere	**19.41**	**0.79**	**2.21**	**95.29**	**96.38**	**53.41**	69.58	21.65	**98.31**	**99.12**
Ours w/o sphere	19.10	**0.78**	2.42	94.79	96.11	52.94	**70.05**	**21.72**	98.11	98.92
(b) Effect of degradation rate θ										
$\theta = 0.0$	**19.46**	**0.79**	**2.19**	**95.40**	96.35	53.39	69.51	**21.75**	**98.33**	**99.13**
$\theta = 0.2$	19.41	**0.79**	2.21	95.29	**96.38**	53.41	69.58	21.65	98.31	99.12
$\theta = 0.5$	19.02	0.78	2.49	95.08	96.24	**53.52**	**69.59**	21.67	98.12	98.92

Table 3: Analysis on number of 3D tracks (left) and patch dimension in Track-the-Noise step (right).

Points	PSNR ↑	SSIM ↑	EPE ↓	CLIP ↑
$N = 1024$	**19.28**	**0.78**	**2.26**	**0.9385**
$N = 512$	19.04	0.78	2.38	0.9373
$N = 256$	18.44	0.77	2.78	0.9347

Patch	PSNR ↑	SSIM ↑	EPE ↓	CLIP ↑
$R = 1.0$	16.95	0.71	5.34	0.9201
$R = 2.0$	19.28	0.78	**2.26**	**0.9385**
$R = 3.0$	**19.40**	**0.79**	2.28	0.9379

4.3 Ablation Studies

3D-sphere noise. We validate 3D-sphere noise sampling in Figure 6 by ablating it and using random noise instead. Without sphere-induced cross-frame noise consistency, the model hallucinates implausible objects and inconsistent layouts, whereas enabling sphere sampling (Move-the-World) yields realistic, coherent motion. Quantitatively on MoveBench single-object (Table 2(a)), sphere noise improves PSNR (19.28 vs 19.10) and EPE (2.26 vs 2.42) and consistently boosts VBench-I2V metrics (consistency, aesthetics, reduced flicker, smoother motion), confirming it is a key component.

Degradation rate. We observe that our 3D-grounded motion noise preserves spatial Gaussian characteristics but lacks temporal Gaussianity. To mitigate this issue, we blend the constructed motion-noise with pure Gaussian noise through alpha blending, using a randomly sampled degradation rate $\theta \in [0, 0.5]$. In Table 2 (b) and Figure 12, a degradation rate of $\theta = 0.2$ provides the best balance between visual quality (PSNR, SSIM, Aesthetics Quality, Imaging Quality) and motion fidelity (EPE, Temporal Flickering, Motion Smoothness). We adopt $\theta = 0.2$ as the default in our experiments.

Point tracks. As we use TAPIP3D [50] to extract 1024 points for each video during preprocessing, and then vary the total of point tracks during training, we provide an experiment to validate the robustness of our method when the number of point tracks are varied. Table 3 (left) and Fig. 13 provides the performance of our model on different number of 3D point tracks. The results show that using 512 and 1024 point tracks yield similar PSNR, SSIM, and CLIP values, while using 1024 leads to more favorable EPEs. Using 256 points yields less competitive results, but the performance remains close to the 512 and 1024 case.

Patch size. We analyze the influence of the patch dimension $(2R - 1)$ when propagating noise using point tracks as described in the Track-the-Noise step (Section 3.1). The results are in Table 3 (right) and Fig. 14. It can be seen that using larger patch size ($R = 2$) yields the best EPE and CLIP while setting $R = 3$ yields the best PSNR and SSIM. Given the minor improvement in PSNR and SSIM, we opt for $R = 2$ as our default setting.

5 Conclusion

We presented UniCaMo, an image-to-video generation framework that unifies camera motion and object motion through a 3D-grounded motion-consistent Gaussian noise representation. By combining noise warping based on 3D point tracks and with 3D-sphere noise sampling, UniCaMo produces motion-consistent noise that guides video diffusion without modifying the underlying model architecture. Our results on MoveBench suggest that structured noise initialization is an effective and efficient way for controllable video generation, enabling independent and joint control over motion.