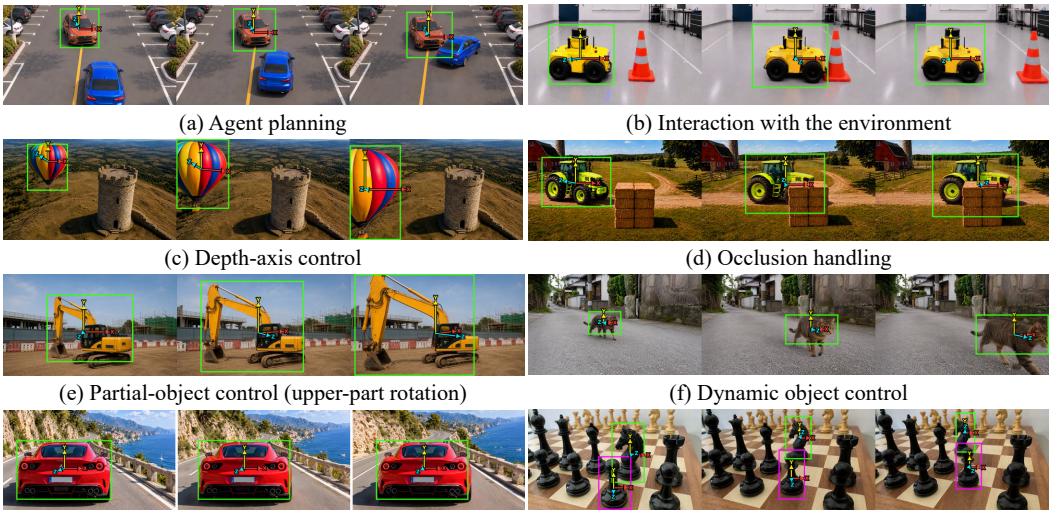
UNISTATE: UNIFYING MEMORY AND INTERVENTION FOR CONTROLLABLE WORLD SIMULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

World models offer increasingly rich environments for exploration, yet intervening in a world while observing the consequences from freely chosen viewpoints remains challenging. We tackle these challenges by introducing **UniState**, a framework for controllable world simulation built on a unified world-state representation. This representation unifies spatial memory and user interventions. Users specify object motion directly within spatial memory, while independently choosing the camera trajectory from which the resulting world state is observed. This unification makes the relationship between the scene and the requested changes explicit, removing ambiguity between object motion and viewpoint change from the control specification. Crucially, specifying an intervention does not require prescribing every resulting movement: the video model can synthesize additional motion within controlled objects and plausible interactions with their surroundings. We develop training and sampling strategies that help generated videos follow the specified object motion and camera trajectory while preserving this capacity to generate responses beyond the prescribed transformations. Experiments demonstrate more accurate joint object and camera control than the existing methods, alongside qualitative evidence that intervention-induced scene changes persist across rollouts.



(a) Agent planning (b) Interaction with the environment

(c) Depth-axis control (d) Occlusion handling

(e) Partial-object control (upper-part rotation) (f) Dynamic object control

(g) World exploration (h) Multi-object control

Figure 1: **Teaser.** UniState, a framework for controllable world simulation, enables coherent and interactive world simulation. We recommend watching our teaser video at `supplementary.html`.

1 INTRODUCTION

World simulation models have been built from video models for their increasing performance of producing high-fidelity future predictions given an initial frame and a description of the events occurring

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.

in the scene (Team Wan et al., 2025; NVIDIA, 2026). They can be used as explorable environments for training agents, especially with camera controls (Wang et al., 2025a; Xing et al., 2025) and object controls (Wang et al., 2024a; Geng et al., 2025).

However, using world models for learning agents is limited to viewpoint exploration (Bar et al., 2025) even though recent world models cover complex real-world environments, and training agents with *physical intervention* to the world is limited to restricted configurations such as grippers with a fixed camera (Du et al., 2023; Yang et al., 2024; Huang et al., 2026). We attribute this gap to recent world models that cover navigation and omit physical intervention.

We suggest that physical intervention beyond navigation is a crucial component for training physical agents, as well as other important requirements including spatio-temporal coherence and free-viewpoint observation. Existing video models exploit persistent spatial memory to support spatio-temporal coherence (Xiao et al., 2025; Yu et al., 2026), camera control to support free-viewpoint observation (He et al., 2025a;b), and dedicated conditions to support object motion (Wu et al., 2025; Wang et al., 2025a). However, none of these support physical intervention in a persistent world with free-viewpoint observation.

We introduce UniState, capable of above three requirements. Briefly, it contains initial spatial memory and user intervention in the same world coordinate system in order to work as a spatio-temporal twin of the world and allow the viewpoint to be separately specified. It removes the ambiguity that the same trajectories can be achieved by either object motion or viewpoint control in previous methods. The initial spatial memory is the latents lifted into 3D from an initial image (Yu et al., 2026). Then users specify motion of a region with a sequence of 6-DoF transformations, enabling 3D object translation and rotation, as illustrated in Fig. 1. Finally, a camera trajectory determines the viewpoints from which UniState is rendered. Last but not least, we introduce sampling and training techniques to reflect all the above variables.

This unified state provides spatial constraints for video generation while leaving the model room to synthesize plausible responses to the specified interventions. For example, an intervention in the surrounding environment can lead the model to generate an agent's evasive maneuver, such as taking a detour to avoid an approaching vehicle (Fig. 1(a)). Controlling a bulldozer can similarly induce motion in nearby objects through contact, without explicitly prescribing their trajectories (Fig. 1(b)). This flexibility also extends to motion within the controlled object: specifying only the translation of a cat's torso allows the model to generate the accompanying leg movements for walking (Fig. 1(f)).

To guarantee the video model to faithfully translate the unified representation into observations from the specified viewpoint, we design both sampling and training strategies for coherent joint object and camera control. Our sampling strategy promotes consistent generation from the specified viewpoint during object interventions, while our training strategy reduces the mismatch between training and inference conditions, helping the video model follow the specified object motion and camera trajectory while retaining its capacity to generate plausible responses.

Experiments show that UniState achieves more accurate joint object and camera control than the evaluated baselines. Qualitative results demonstrate manipulation of whole objects and selected parts with consistent depth ordering and occlusion, alongside motion and interactions beyond the prescribed transformations. Ablation studies support the benefits of the proposed training and sampling strategies for control accuracy.

Our key contributions are:

- With UniState, we show that unifying spatial memory and user interventions in a unified world-state representation allows object motion and camera trajectory to be specified independently.
- We show that using the initial state as a reference during sampling and reducing the mismatch between training and inference conditions improve joint object and camera control.

2 Related work

Video world models as interactive simulators. Video world models simulate the consequences of actions for planning and policy learning, using generated videos as plans for control (Du et al.,

Table 1: Comparison of spatial memory and object-intervention representations. UniState is the only method that unifies spatial memory and user intervention within a unified representation and world coordinate system. Detailed capability and intervention-representation comparisons are provided in Table 5 and Table 6.

Method	Spatial memory	User intervention	Unified repr.
VerseCrafter	RGB point cloud	3D Gaussians	✘
WorldDirector	Video context tokens	2D box trajectories	✘
MotionCanvas	Image latent	2D boxes / point tracks	✘
SymphoMotion	Point-cloud render	3D point trajectories	✘
Real2SAM2Real	Video latent	3D mesh	✘
Ours	UniState		✔

2023) or as interactive environments for policy training (Yang et al., 2024). Action-conditioned video models further leverage pretrained video priors or learn from latent actions and recorded controls for interactive simulation (Rigter et al., 2025; Huang et al., 2026; Bruce et al., 2024; Alonso et al., 2024; Valevski et al., 2025; Li et al., 2025a; He et al., 2026). Recent methods extend exploration across diverse scenes, using context management and accelerated generation to sustain interaction (Mao et al., 2025; Sun et al., 2026; Team et al., 2026; Ball et al., 2025). Complementary approaches use predicted observations for navigation planning (Bar et al., 2025), incorporate geometric prediction (Zhu et al., 2025; Chen et al., 2025), or ground exploration in geographically indexed street views (Seo et al., 2026). Instruction-driven behaviors and object-directed actions further broaden interaction beyond navigation (Tang et al., 2026; Gao et al., 2026). UniState extends world exploration with time-varying 6-DoF interventions on selected object regions with the unified representation and world coordinate system as persistent memory, while allowing the camera trajectory to be specified independently.

Object control and camera control. Methods for controlling motion in video generation specify motion through point trajectories (Wu et al., 2025; Zhang et al., 2025b; Geng et al., 2025; Wang et al., 2025a) or bounding boxes (Wang et al., 2024a); joint object and camera control further requires coordinating these motion signals with viewpoint conditions (Wang et al., 2024b; Xing et al., 2025), often using depth, pose, or 3D guidance (Wang et al., 2025b; Ji et al., 2025; Gu et al., 2025; Cao et al., 2025; Wang et al., 2026a; Zhang et al., 2026). Since viewpoint change and object motion both induce image-space displacement, recent methods disentangle them through cross-view transfer (Liu et al., 2026) or dynamic reprojection of world-anchored trajectories (Gu et al., 2026). Methods using explicit 3D representations condition video generation on Gaussians expressed in a shared coordinate frame (Zheng et al., 2026) or editable foreground caches (Wu et al., 2026), while appearance-aware methods transport latent features along point tracks (Chu et al., 2025; Gu et al., 2026). As summarized in Table 1, the compared methods encode spatial memory and user intervention in separate representations. This separation leaves generation susceptible to background drift and ambiguity between object motion and viewpoint change. UniState grounds both in a shared appearance-bearing 3D representation and world coordinate system, making their spatial relationship explicit.

Persistent and editable spatial memory. While camera conditioning controls the viewpoint (He et al., 2025a;b), scene consistency can be maintained by retrieving spatially relevant observations (Xiao et al., 2025; Yu et al., 2025; Sun et al., 2026) or compressing historical context (Zhang et al., 2025a; Oshima et al., 2026; Hong et al., 2025). Retrieval and compression retain scene evidence but do not themselves define geometric operations on entities. Geometry-based approaches render, warp, or retrieve scene information using explicit spatial memories (Ren et al., 2025; Wang et al., 2026c; Zhao et al., 2025; Li et al., 2025b), with some excluding dynamic objects from persistent memory (Wang et al., 2026c). Other memory-based approaches maintain out-of-view dynamics (Chen et al., 2026; Duan et al., 2026) or combine coordinated 3D object and camera trajectories with persistent dynamic object memory (Wang et al., 2026a). Spatial patch memories further support relocation, duplication, and deletion, but focus on scene composition rather than prescribed instance-level motion over time (Yu et al., 2026). Explicit spatial representations also support editing through evolving voxel states, but validation of this approach remains limited to game environments (Garcin

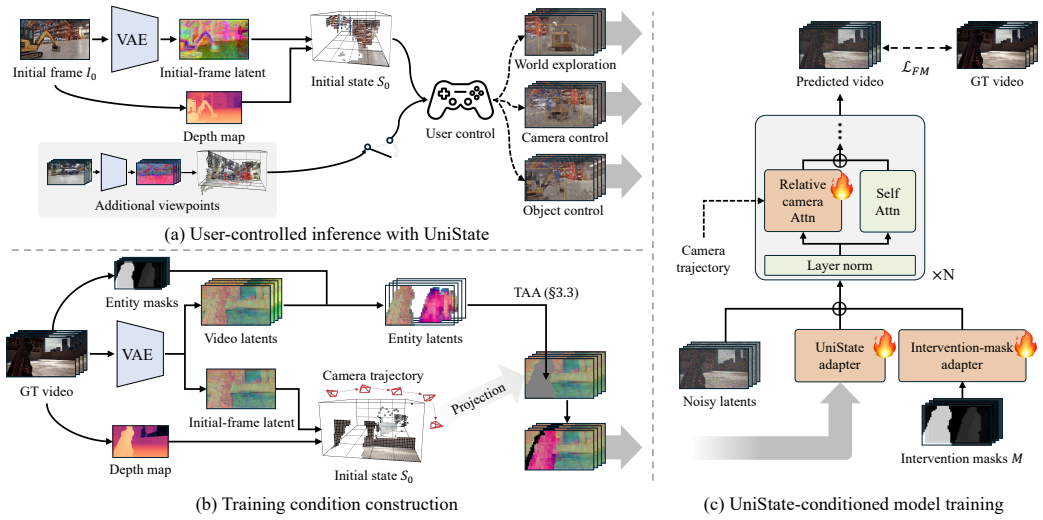


Figure 2: **Overview of UniState.** (a) Source observations are lifted into a unified state. Users specify interventions and a camera trajectory to construct the inference conditions. (b) During training, entity latents from ground-truth videos are combined with initial-state renders by masked compositing. (c) Only the conditioning modules marked with flames are trained. $\mathcal{L}_{\text{FM}}$ denotes the flow-matching loss between predicted and target latent velocities.

et al., 2026). UniState applies time-varying 6-DoF transformations to whole or partial object regions within shared 3D spatial memory, while the video prior synthesizes dynamics beyond the prescribed rigid motion.

3 METHOD

UniState generates a video from an initial image, user-specified interventions, and a camera trajectory. We lift initial-frame latent features into 3D to construct spatial memory, then express user interventions by transforming the geometry of selected regions within this memory. The transformed entities and unchanged scene context thus remain in a single set of 3D latents, which we call a unified *world-state representation*. Rendering this state along an independently specified camera trajectory provides the conditions for video generation. Fig. 2 summarizes the inference and training pipelines.

We first describe how UniState constructs, manipulates, and renders this unified state (Section 3.1). We then introduce Scene-Anchored Joint Guidance (Section 3.2) and a training strategy (Section 3.3) that help the video model follow the resulting conditions.

3.1 UNIFIED WORLD-STATE REPRESENTATION

Constructing spatial memory. We encode the initial image into VAE latent features and lift them into 3D using the source camera and estimated depth (Wang et al., 2026b). Each resulting 3D latent pair has geometry with its source feature; together, these latents form the initial state S_0. Additional views of the same initial scene can extend S_0 by lifting and registering their latent features in the same world coordinate system. Geometric construction details are provided in Appendix A.1.

Expressing interventions within the state. Source-view segmentation masks, provided by the user or obtained with SAM 2 (Ravi et al., 2025), identify the latent index set $\mathcal{I}_k$ for each entity k, which may be a whole object or a selected part. For each entity, the user specifies rigid[1] transformations $\{G_{t,k} \in \mathrm{SE}(3)\}_{t=1}^{T}$ in the shared world coordinate system over T subsequent frames. Let $X_{0,i}$

[1]Part-wise rigid transformations can approximate non-rigid motion.

denote the geometry of the i-th 3D latent in the initial state S_0. At each time t, its geometry is

$$X_{t,i} = \begin{cases} G_{t,k} X_{0,i}, & i \in \mathcal{I}_k, \\ X_{0,i}, & \text{otherwise.} \end{cases} \tag{1}$$

We construct each intervened state S_t directly from S_0 using these transformed geometries, while preserving source latent features and leaving unselected regions unchanged.

Rendering the state along the camera trajectory. Given an independently specified camera trajectory $C = \{C_t\}_{t=1}^{T}$, a visibility-aware renderer $\mathcal{R}$ projects the intervened state into each camera view: $(W_t, M_t) = \mathcal{R}(S_t; C_t)$. Here, W_t contains the projected source latent features, and M_t marks the visible regions of selected entities. Visibility is resolved against the complete scene; rendering details are provided in Appendix A.1.

The camera-conditioning module, UniState adapter, and intervention-mask adapter process the relative-ray camera condition C, latent condition $W = \{W_t\}_{t=1}^{T}$, and intervention masks $M = \{M_t\}_{t=1}^{T}$. Their outputs are added to the video tokens. These partial observations of the unified state guide generation while the video model synthesizes the remaining scene content and motion.

Interaction and state updates. User interventions specify the motion of the controlled entity, while the video model synthesizes plausible responses in the surrounding scene. Conditioning scene regions involved in an interaction on their previous configuration can constrain these responses. We therefore temporarily exclude these regions from the memory condition while retaining the prescribed control over the entity. This allows nearby objects to move in response to contact without explicitly prescribing their trajectories.

After the interaction, we obtain updated geometry and latent features for the affected regions from generated observations, register the geometry to the shared world coordinate system, and update the corresponding state regions. The updated state serves as S_0 for the next rollout, allowing subsequent viewpoint changes and interventions to build on the resulting scene configuration.

3.2 Scene-Anchored Joint Guidance

Our conditional video model f_θ is trained with a flow-matching objective and predicts latent velocities during sampling.

Although the unified state specifies interventions independently of camera viewpoint, the generated video may still entangle their effects, causing background motion to deviate from the prescribed camera trajectory. Our key observation is that this ambiguity can be removed by constructing a camera-only reference: if we render the initial state S_0 under the same camera trajectory but without applying any intervention, the resulting condition should contain scene changes induced by viewpoint change alone. We therefore use the velocity conditioned on the intervention for the intervened region and use the velocity conditioned on the initial state for the background. We spatially route between these two predictions and blend the result with the full joint prediction. We call this procedure Scene-Anchored Joint Guidance (SAJG).

Scene-anchor and joint predictions. Using the same camera trajectory, we render the initial state S_0 without interventions to obtain the scene-anchor condition W_s, and the intervened states $\{S_t\}_{t=1}^{T}$ to obtain the joint condition W_j, corresponding to W in Section 3.1.

Using the same noisy video latent, denoising timestep, camera condition, source observation, and text prompt, we evaluate f_θ under the scene-anchor and joint conditions to obtain the respective velocity predictions V_s and V_j. Both predictions use classifier-free guidance; implementation details are provided in Appendix A.2.

Spatial routing. The scene-anchor prediction provides a reference for camera-induced changes in the surrounding scene, while the joint prediction incorporates the requested interventions. We therefore use the joint prediction in regions affected by these transformations, including the entities' target locations and the regions they vacate. A soft scene mask G_s excludes these projected source and target regions, combining the predictions as

$$V_{\text{routed}} = G_s \odot V_s + (1 - G_s) \odot V_j, \tag{2}$$

where $\odot$ denotes element-wise multiplication. Mask construction details are provided in Appendix A.2.

Joint prediction anchoring. To retain the joint prediction's contribution throughout the video, including scene regions assigned to the scene-anchor prediction, we blend the spatially routed prediction with the full joint prediction:

$$V_{\text{final}} = \alpha V_j + (1 - \alpha) V_{\text{routed}}, \qquad \alpha \in [0, 1]. \tag{3}$$

Here, α controls the weight of the full joint prediction in the final blend, ensuring a minimum joint contribution of α at every spatio-temporal location. Setting $\alpha = 0$ uses the spatially routed prediction alone, whereas $\alpha = 1$ uses only the joint prediction. We use $\alpha = 0.25$ unless stated otherwise. At each denoising step, V_{final} is passed to the flow-matching solver.

3.3 Training

Training conditions. We train on videos paired with camera conditions C, latent conditions W, and intervention masks M derived from per-frame object masks. We construct W by combining initial-state renders with ground-truth entity latents. The latter preserve alignment with the target video, avoiding misalignment and floating artifacts caused by transporting source features with imperfect geometry (see Fig. 10 in the appendix).

We render the initial state from each target camera and zero out the selected entities' projected source regions. We then insert ground-truth entity latents at their target locations, retaining the initial-state render elsewhere. Vacated source regions remain zero unless covered by a target entity. The exact compositing formulation is provided in Appendix A.3.

Transport-Aligned Augmentation. The composited conditions introduce a training–inference mismatch: training uses entity latents from each ground-truth target frame, whereas inference transports source latents through the transformed geometry. To reduce this mismatch, we apply Transport-Aligned Augmentation (TAA), which spatially coarsens future-frame entity latents while preserving their target support. TAA modifies only the latent condition; the intervention masks and ground-truth training targets remain unchanged. The augmentation procedure is detailed in Appendix A.3.

Curriculum training. We train UniState in two stages using a flow-matching objective $\mathcal{L}_{\text{FM}}$, keeping the pretrained video backbone, VAE, and text encoder frozen. The first stage trains the camera-conditioning module and UniState adapter on DL3DV (Ling et al., 2024). Starting from these weights, the second stage adds the intervention-mask adapter and jointly trains all three modules on HM-World(Chen et al., 2026) using the composited conditions and TAA described above.

4 Experiments

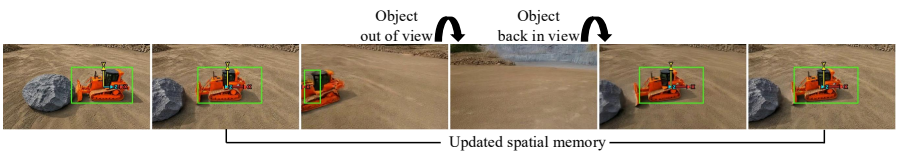
4.1 Implementation details

We use a pretrained Cosmos3-Nano I2V model (NVIDIA, 2026) and follow the two-stage curriculum in Section 3.3, with 20K optimization steps on DL3DV followed by 10K on HM-World. We generate 81-frame videos at 832×480 resolution and 16 fps. Optimization settings and caption construction are detailed in Appendix A.3.

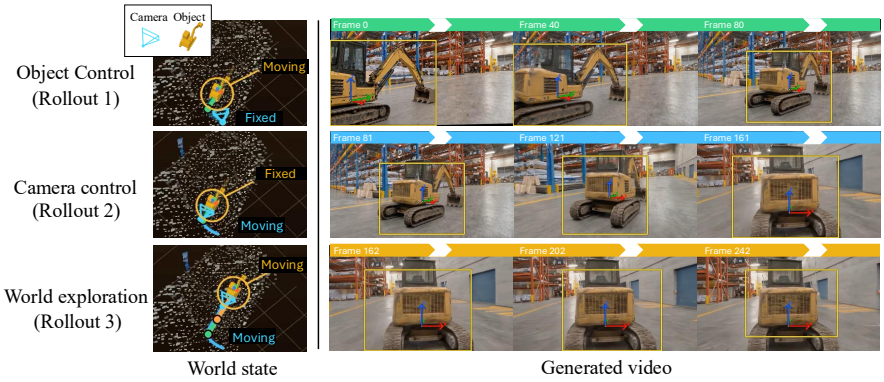
4.2 Evaluation

We evaluate all methods on the same 277 scenarios spanning diverse camera trajectories and user interventions. Each method generates two videos per scenario with independent random seeds (554 videos total), using an identical evaluation protocol and averaging metrics across the two seeds.

Joint control accuracy We introduce World-Control Projection Error (WCPE) to evaluate object control specified in world coordinates. Because video models produce 2D observations, an object's screen-space position alone does not establish whether it follows the specified motion in the world. We therefore assess object motion relative to the surrounding scene, while checking that the scene



(a) State persistence after intervention and revisitation



(b) Persistent world control

Figure 3: **State persistence and control across rollouts.** (a) A rock displaced by the controlled bulldozer remains in its updated position after the camera leaves and revisits the region. (b) Three successive rollouts demonstrate object control with a fixed camera, camera control with a fixed object pose, and joint object and camera control, while carrying the world state forward. The left column visualizes the unified state and prescribed controls; the right columns show generated frames. Colored boxes identify controlled entities, and coordinate axes indicate their prescribed poses.

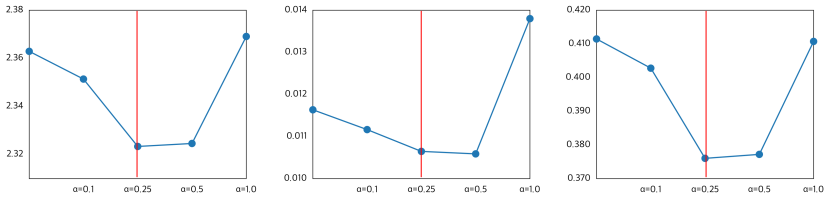
remains aligned with the specified camera trajectory. WCPE implements this assessment as a 2D proxy, using target projections to evaluate object–background relative positions and background alignment. Requiring both prevents correct screen-space object placement alone from yielding a low error when the surrounding scene is misaligned. The metric formulation and evaluation protocol are detailed in Section A.4.

Camera-control accuracy We evaluate camera-control accuracy by comparing the camera trajectory recovered from each generated video with its target trajectory. We report Cam-T, the mean translation error between the recovered and target camera positions, and Cam-R, the mean geodesic rotation error between their orientations, measured in degrees. Lower values for both metrics indicate more faithful camera control.

4.3 Qualitative Results

Diverse control through user interventions. Fig. 1(c) and (d) demonstrate depth-axis control and occlusion handling consistent with scene geometry. Interventions can also target selected object parts: in Fig. 1(e), the user rotates the excavator's upper structure independently of its tracks, specifying the rotation direction and angular speed. For non-rigid objects, partial control allows the video model to complete unspecified articulation. In Fig. 1(f), the user specifies only the cat's torso trajectory, while the model generates accompanying limb motion and maintains a coherent body configuration. Fig. 1(g) and (h) further demonstrate camera-controlled world exploration and simultaneous control of multiple objects, respectively.

State persistence and control across rollouts. In Fig. 3, the controlled bulldozer displaces a passive rock. After the resulting state is incorporated into memory, the camera moves away until the affected region is out of view, then returns. The rock retains its displaced position upon revisitation, providing qualitative evidence of state persistence across rollouts. Fig. 3(b) demonstrates continued control of the same world across three successive rollouts. The user first moves the excavator with the camera fixed, then changes the viewpoint while preserving the excavator's updated world-space



(a) WCPE (%) ↓ (b) Cam-T ↓ (c) Cam-R (°) ↓

Figure 4: Ablation results for Scene-Anchored Joint Guidance. The red vertical line denotes the selected setting $\alpha = 0.25$.

Table 2: Ablation study on Transport-Aligned Augmentation (TAA). † denotes using transported initial-frame latents as the entity condition during training.

Training condition	WCPE (%) ↓	Cam-T ↓	Cam-R (°) ↓
Direct source-transport†	2.857	0.0115	0.405
GT condition w/o TAA	2.574	0.0114	0.412
GT condition + TAA	**2.323**	**0.0106**	**0.376**

pose, and finally controls both the object and camera. Each rollout continues from the preceding state, allowing the user to manipulate an entity, observe the resulting configuration from another viewpoint, and intervene again. This separates control of the entity's world state from control of the observation viewpoint.

4.4 Ablation Study

Scene-Anchored Joint Guidance. Fig. 4 shows that intermediate values of α outperform both spatial routing alone ($\alpha = 0$) and the joint prediction alone ($\alpha = 1$). We select $\alpha = 0.25$, which achieves the lowest WCPE and Cam-R and nearly the lowest Cam-T. These results support combining a scene-anchor correction with a global joint contribution.

Transport-Aligned Augmentation. We compare three ways of constructing entity latent conditions during training: transporting initial-frame latents using estimated geometry, using aligned ground-truth entity latents from each target frame, and applying TAA to those ground-truth latents. All variants share training data, initialization, optimization budgets, inference conditions, and SAJG settings. As shown in Table 2, replacing transported initial-frame latents with ground-truth latents reduces WCPE and Cam-T, but slightly increases Cam-R. Applying TAA further reduces all three errors, supporting the combination of spatially aligned training conditions and augmentation that reduces their latent mismatch with inference-time transport.

Curriculum training. Staged training outperforms interleaved training under matched data exposure and computation; details are provided in Appendix B.1.

4.5 Comparisons

We compare our method with VerseCrafter (Zheng et al., 2026), WorldDirector (Wang et al., 2026a), MotionCanvas (Xing et al., 2025), SymphoMotion (Zhang et al., 2026), and Real2SAM2Real (Wu et al., 2026). We evaluate joint object and camera control using WCPE, Cam-T, and Cam-R. As shown in Table 3, our method achieves the lowest WCPE, Cam-T, and Cam-R, demonstrating accurate joint control of object motion and camera viewpoint. Qualitative comparisons in Fig. 5 further show these failure modes. In this example, the user keeps the camera static while translating and rotating the ball laterally behind the alphabet blocks. VerseCrafter and Real2SAM2Real often absorb object motion into viewpoint change, resulting in undesired viewpoint change despite the specified static camera condition. This makes the ball appear to move with the camera rather than independently in the scene. WorldDirector and MotionCanvas preserve the static camera more reliably, but

Table 3: Comparison with competing methods on joint object and camera control. **Bold** denotes the best score, and underline denotes the second-best score for each metric.

Method	WCPE (%) ↓	Cam-T ↓	Cam-R (°) ↓
VerseCrafter	3.039	0.0540	0.385
WorldDirector	2.487	0.0844	0.571
SymphoMotion	2.822	0.0627	0.526
MotionCanvas	2.425	0.0378	0.428
Real2SAM2Real	3.061	0.104	0.608
Ours	**2.323**	**0.0106**	**0.376**

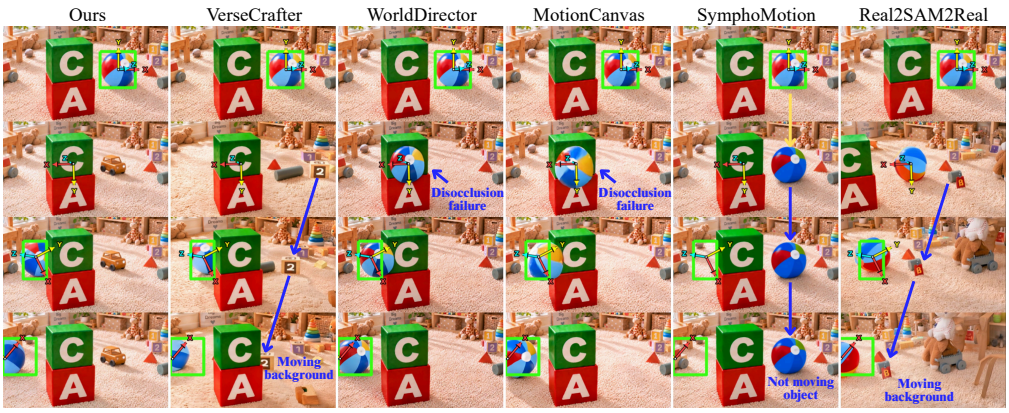


Figure 5: **Object translation and rotation control under occlusion.** The ball translates from right to left and rotates behind the alphabet blocks. Green boxes and XYZ axes indicate its target position and orientation. UniState preserves object motion, occlusion, and background stability. See `supplementary.html` for the full sequence.

fail to reproduce the correct depth ordering and rotation of the ball. In contrast, our method faithfully follows the intended joint object and camera control, capturing object translation and rotation under a static camera while handling occlusion. See Section C for additional qualitative comparisons. An additional comparison with MosaicMem is provided in Table 7.

5 Conclusion

UniState unifies spatial memory and user interventions in a shared world coordinate system for controllable world simulation. This unified world-state representation allows object motion and camera trajectory to be specified independently. It provides spatial constraints for video generation while preserving the video model's capacity to synthesize plausible responses, so users can specify intended changes without prescribing every resulting motion or interaction. Quantitative comparisons and ablations support our training and sampling strategies for joint object and camera control across diverse camera trajectories and user interventions. Revisitation results further show that generated interaction outcomes can be incorporated into spatial memory and retained across rollouts. Together, these findings support physical intervention in a persistent world with free-viewpoint observation.

Limitations Interaction effects rely on the video model's generative prior: although controlled object states can be updated outside the camera view, their interactions with surrounding objects cannot be explicitly simulated while out of view. Spatial resolution limits control over small objects, while temporal resolution limits fast motion. The model may also fail to follow physically implausible interventions, such as moving a solid object through a solid wall.

AI USE STATEMENT

We used generative AI tools to assist in implementing evaluation code for comparing our method with competing methods on the same test set, identifying relevant related work, and improving the readability of the manuscript. We also used Qwen2.5-VL (Bai et al., 2025) to generate captions for training videos, as described in Appendix A.3. AI tools were not used to design the experiments. The authors take full responsibility for the final code, results, references, and manuscript.

REFERENCES

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 58757–58791. Curran Associates, Inc., 2024. doi: 10.52202/079017-1873. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/6bdde0373d53d4a501249547084bed43-Paper-Conference.pdf.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Philip J. Ball, Jakob Bauer, Frank Belletti, Bethanie Brownfield, Ariel Ephrat, Shlomi Fruchter, Agrim Gupta, Kristian Holsheimer, Aleksander Holynski, Jiri Hron, Christos Kaplanis, Marjorie Limont, Matt McGill, Yanko Oliveira, Jack Parker-Holder, Frank Perbet, Guy Scully, Jeremy Shar, Stephen Spencer, Omer Tov, Ruben Villegas, Emma Wang, Jessica Yung, Cip Baetu, Jordi Berbel, David Bridson, Jake Bruce, Gavin Buttimore, Sarah Chakera, Bilva Chandra, Paul Collins, Alex Cullum, Bogdan Damoc, Vibha Dasagi, Maxime Gazeau, Charles Gbadamosi, Woohyun Han, Ed Hirst, Ashyana Kachra, Lucie Kerley, Kristian Kjems, Eva Knoepfel, Vika Koriakin, Jessica Lo, Cong Lu, Zeb Mehring, Alex Moufarek, Henna Nandwani, Valeria Oliveira, Fabio Pardo, Jane Park, Andrew Pierson, Ben Poole, Helen Ran, Tim Salimans, Manuel Sanchez, Igor Saprykin, Amy Shen, Sailesh Sidhwani, Duncan Smith, Joe Stanton, Hamish Tomlinson, Dimple Vijaykumar, Luyu Wang, Piers Wingfield, Nat Wong, Keyang Xu, Christopher Yew, Nick Young, Vadim Zubov, Douglas Eck, Dumitru Erhan, Koray Kavukcuoglu, Demis Hassabis, Zoubin Gharamani, Raia Hadsell, Aäron van den Oord, Inbar Mosseri, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 3: A new frontier for world models, 2025.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 15791–15801, June 2025.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C.Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando De Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4603–4623. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/bruce24a.html.

Chenjie Cao, Jingkai Zhou, Shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, SA Conference Papers '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400721373. doi: 10.1145/3757377.3763842. URL https://doi.org/10.1145/3757377.3763842.

Junyi Chen, Haoyi Zhu, Xianglong He, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Zhoujie Fu, Jiangmiao Pang, and Tong He. Deepverse: 4d autoregressive video generation as a world model, 2025. URL https://arxiv.org/abs/2506.01103.

Kaijin Chen, Dingkang Liang, Xin Zhou, Yikang Ding, Xiaoqiang Liu, Pengfei Wan, and Xiang Bai. Out of sight but not out of mind: Hybrid memory for dynamic video world models, 2026. URL https://arxiv.org/abs/2603.25716.

Ruihang Chu, Yefei He, Zhekai Chen, Shiwei Zhang, Xiaogang Xu, bin xia, Dingdong WANG, Hongwei Yi, Xihui Liu, Hengshuang Zhao, Yu Liu, Yingya Zhang, and Yujiu Yang. Wan-move: Motion-controllable video generation via latent trajectory guidance. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 404–432. Curran Associates, Inc., 2025. doi: 10.52202/085713-0014. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/00b67df24009747e8bbed4c2c6f9c825-Paper-Conference.pdf.

Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 9156–9172. Curran Associates, Inc., 2023. doi: 10.52202/075280-0403. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1d5b9233ad716a43be5c0d3023cb82d0-Paper-Conference.pdf.

Zicheng Duan, Jiatong Xia, Zeyu Zhang, Wenbo Zhang, Gengze Zhou, Chenhui Gou, Yefei He, Feng Chen, Xinyu Zhang, and Lingqiao Liu. Liveworld: Simulating out-of-sight dynamics in generative video world models, 2026. URL https://arxiv.org/abs/2603.07145.

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, Ka Leong Cheng, Haojie Zhang, Jian Gao, Tianrui Feng, Yuzheng Liu, Yao Yao, Yinghao Xu, Xing Zhu, Yujun Shen, and Hao Ouyang. Infinite worlds with versatile interactions, 2026. URL https://arxiv.org/abs/2607.07534.

Samuel Garcin, Thomas Walker, Steven McDonagh, Tim Pearce, Hakan Bilen, Tianyu He, Kaixin Wang, and Jiang Bian. Beyond pixel histories: World models with persistent 3d state. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=yyitz0hJqK.

Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, Chen Sun, Oliver Wang, Andrew Owens, and Deqing Sun. Motion prompting: Controlling video generation with motion trajectories. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 1–12, June 2025.

Bohai Gu, Taiyi Wu, Yueyang Yuan, Jian Liu, Xiaocheng Lu, Dazhao Du, Jie Zhang, Jinxiang Lai, Shuai Yang, Xiaotong Zhao, Alan Zhao, and Song Guo. Worldcraft: From camera navigation to object manipulation in interactive video world models, 2026. URL https://arxiv.org/abs/2605.25077.

Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, Wenping Wang, and Yuan Liu. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, SIGGRAPH Conference Papers '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715402. doi: 10.1145/3721238.3730607. URL https://doi.org/10.1145/3721238.3730607.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for video diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=Z4evOUYrk7.

Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13416–13426, October 2025b.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, Baixin Xu, Hao-Xiang Guo, Kaixiong Gong, Size Wu, Wei Li, Xuchen Song, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-game 2.0: An open-source real-time and streaming interactive world model, 2026. URL https://arxiv.org/abs/2508.13009.

Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, Sai Bi, Yannick Hold-Geoffroy, Mike Roberts, Matthew Fisher, Eli Shechtman, Kalyan Sunkavalli, Feng Liu, Zhengqi Li, and Hao Tan. Relic: Interactive video world model with long-horizon memory, 2025. URL https://arxiv.org/abs/2512.04040.

Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2world: Crafting video diffusion models to interactive world models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=pFyzqbUiF9.

Longbin Ji, Lei Zhong, Pengfei Wei, and Changjian Li. Posetraj: Pose-aware trajectory control in video diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 22776–22785, June 2025.

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labeling real videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–10. IEEE, 2025.

Jiaqi Li, Junshu Tang, Zhiyong Xu, Longhuang Wu, Yuan Zhou, Shuai Shao, Tianbao Yu, Zhiguo Cao, and Qinglin Lu. Hunyuan-gamecraft: High-dynamic interactive game video generation with hybrid history condition, 2025a. URL https://arxiv.org/abs/2506.17201.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25690–25699, October 2025b.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22160–22169, June 2024.

Shaowei Liu, Xuanchi Ren, Tianchang Shen, Huan Ling, Saurabh Gupta, Shenlong Wang, Sanja Fidler, and Jun Gao. Moright: Motion control done right, 2026. URL https://arxiv.org/abs/2604.07348.

Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, Tong He, Jiangmiao Pang, Yu Qiao, and Kaipeng Zhang. Yume-1.5: A text-controlled interactive world generation model, 2025. URL https://arxiv.org/abs/2512.22096.

NVIDIA. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026. URL https://research.nvidia.com/labs/cosmos-lab/cosmos3/technical-report.pdf.

Yuta Oshima, Yusuke Iwasawa, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. Worldpack: Dynamic frame compression for long-context video world modeling. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=zJuiG3PiNJ.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev

Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Ha6RTeWMd0.

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 Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 6121–6132, June 2025.

Marc Rigter, Tarun Gupta, Agrin Hilmkil, and Chao Ma. AVID: Adapting video diffusion models to world models. In *Reinforcement Learning Conference*, 2025. URL https://openreview.net/forum?id=C18kcGeqAW.

Junyoung Seo, Hyunwook Choi, Minkyung Kwon, Jinhyeok Choi, Siyoon Jin, Gayoung Lee, Junho Kim, JoungBin Lee, Geonmo Gu, Dongyoon Han, Sangdoo Yun, Seungryong Kim, and Jin-Hwa Kim. Grounding world simulation models in a real-world metropolis, 2026. URL https://arxiv.org/abs/2603.15583.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling, 2026. URL https://arxiv.org/abs/2512.14614.

Junshu Tang, Jiacheng Liu, Jiaqi Li, Longhuang Wu, Haoyu Yang, Penghao Zhao, Siruis Gong, Xiang Yuan, Shuai Shao, Linfeng Zhang, and Qinglin Lu. Hunyuan-gamecraft-2: Instruction-following interactive game world model, 2026. URL https://arxiv.org/abs/2511.23429.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, Yihang Chen, Jie Liu, Yansong Cheng, Yao Yao, Jiayi Zhu, Yihao Meng, Kecheng Zheng, Qingyan Bai, Jingye Chen, Zehong Shen, Yue Yu, Xing Zhu, Yujun Shen, and Hao Ouyang. Advancing open-source world models, 2026. URL https://arxiv.org/abs/2601.20540.

Team Wan et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=P8pqeEkn1H.

Angtian Wang, Haibin Huang, Jacob Zhiyuan Fang, Yiding Yang, and Chongyang Ma. Ati: Any trajectory instruction for controllable video generation, 2025a. URL https://arxiv.org/abs/2505.22944.

Hanlin Wang, Hao Ouyang, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Qifeng Chen, Yujun Shen, and Limin Wang. Levitor: 3d trajectory oriented image-to-video synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 12490–12500, June 2025b.

Hanlin Wang, Hao Ouyang, Qiuyu Wang, Wen Wang, Qingyan Bai, Ka Leong Cheng, Yue Yu, Yixuan Li, Yihao Meng, Zichen Liu, Yanhong Zeng, Yujun Shen, and Qifeng Chen. WorldDirector: Building controllable world simulators with persistent dynamic memory, 2026a. URL https://arxiv.org/abs/2607.02517.

Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. Boximator: Generating rich and controllable motions for video synthesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52274–52289. PMLR, 21–27 Jul 2024a. URL https://proceedings.mlr.press/v235/wang24cr.html.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *Advances in Neural Information Processing Systems*, 38:35928–35959, 2026b.

Weijie Wang, Haoyu Zhao, Yifan Yang, Feng Chen, Zeyu Zhang, Yefei He, Zicheng Duan, Donny Y. Chen, Yuqing Yang, and Bohan Zhuang. Latent spatial memory for video world models, 2026c. URL https://arxiv.org/abs/2606.09828.

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*, SIGGRAPH '24, New York, NY, USA, 2024b. Association for Computing Machinery. ISBN 9798400705250. doi: 10.1145/3641519.3657518. URL https://doi.org/10.1145/3641519.3657518.

Jiayi Wu, Haoming Cai, Cornelia Fermuller, Christopher Metzler, and Yiannis Aloimonos. Real2SAM2Real: Generative 3d caches as complementary context for video diffusion, 2026. URL https://arxiv.org/abs/2606.00299.

Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. Draganything: Motion control for anything using entity representation. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 331–348, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72670-5.

Zeqi Xiao, Yushi LAN, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 49632–49652. Curran Associates, Inc., 2025. doi: 10.52202/085713-1659. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/470629a47e2d65ce0606c40055df5d26-Paper-Conference.pdf.

Jinbo Xing, Long Mai, Cusuh Ham, Jiahui Huang, Aniruddha Mahapatra, Chi-Wing Fu, Tien-Tsin Wong, and Feng Liu. MotionCanvas: Cinematic shot design with controllable image-to-video generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, SIGGRAPH Conference Papers '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715402. doi: 10.1145/3721238.3730604. URL https://doi.org/10.1145/3721238.3730604.

Sherry Yang, Yilun Du, Seyed Kamyar Seyed Ghasemipour, Jonathan Tompson, Leslie Pack Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=sFyTZEqmUY.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, SA Conference Papers '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400721373. doi: 10.1145/3757377.3763833. URL https://doi.org/10.1145/3757377.3763833.

Wei Yu, Runjia Qian, Yumeng Li, Liquan Wang, Songheng Yin, Sri Siddarth Chakaravarthy P, Dennis Anthony, Yang Ye, Yidi Li, Weiwei Wan, and Animesh Garg. MosaicMem: Hybrid spatial memory for controllable video world models, 2026. URL https://arxiv.org/abs/2603.17117.

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 (CVPR)*, pp. 11127–11137, June 2026.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 30546–30566. Curran Associates, Inc., 2025a. doi: 10.52202/085713-1024. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/2bde8fef08f7ebe42b584266cbcfc909-Paper-Conference.pdf.

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 (CVPR)*, pp. 2063–2073, June 2025b.

Jinjing Zhao, Fangyun Wei, Zhening Liu, Hongyang Zhang, Chang Xu, and Yan Lu. Spatia: Video generation with updatable spatial memory, 2025. URL https://arxiv.org/abs/2512.15716.

Sixiao Zheng, Minghao Yin, Wenbo Hu, Xiaoyu Li, Ying Shan, and Yanwei Fu. VerseCrafter: Dynamic realistic video world model with 4d geometric control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40277–40290, June 2026.

Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. Aether: Geometric-aware unified world modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8535–8546, October 2025.

A IMPLEMENTATION DETAILS

A.1 UNISTATE CONSTRUCTION

Geometry and feature alignment. Given an initial image I_0 and its camera C_0, we obtain latents $z_0 = \mathcal{E}(I_0)$ using a VAE encoder and estimate depth D_0 using an off-the-shelf estimator (Wang et al., 2026b). Using D_0 and C_0, we associate each feature $z_{0,i}$ with 3D geometry $X_{0,i}$ in the world coordinate system defined by C_0. These geometry–feature pairs form the initial state $S_0 = \{\mathcal{P}_{0,i}\}_{i=1}^{N}$, where $\mathcal{P}_{0,i} = (X_{0,i}, z_{0,i})$ denotes a 3D latent. Additional views of the same initial scene can extend S_0 by lifting and registering their latent features into this coordinate system.

Let $\{\Omega_i\}_{i=1}^{N}$ denote the initial image patches. We construct the geometric support of patch i as

$$X_{0,i} = \Pi_{C_0}^{-1}\left(\Omega_i, D_0[\Omega_i]\right),$$

where $\Pi_{C_0}^{-1}$ denotes unprojection through the source camera. The implementation approximates the resulting local geometry with an oriented planar primitive. Its appearance grid is sampled from the aligned latent coordinates U_i.

Intervention. Entity geometry is transformed according to Eq. (1). At each time t, we construct $S_t = \{\mathcal{P}_{t,i}\}_{i=1}^{N}$ directly from S_0, where $\mathcal{P}_{t,i} = (X_{t,i}, z_{0,i})$. Thus, each transformed geometry remains paired with its source latent feature.

Patch selection. Given an entity mask S_k, patch i is assigned to entity k when

$$\frac{|\Omega_i \cap S_k|}{|\Omega_i|} > \tau_{\text{seg}}.$$

We store a control indicator $\alpha_i = \mathbb{1}[i \in \bigcup_k \mathcal{I}_k]$ separately from the transform. This indicator remains active when the selected entity holds a constant pose.

Back-face rejection. Let $n_{i,t}$ and $\bar{x}_{t,i}$ denote the transformed surface normal and patch center. Given target camera center o_t, patch i is rendered when

$$\langle n_{i,t}, o_t - \bar{x}_{t,i} \rangle > 0.$$

Patches at an angle of 90° or greater relative to the front-facing viewing direction are rejected.

Intervention support. We first rasterize every selected and front-facing patch:

$$\widetilde{M}_t = \mathcal{R}_{\text{support}}\left(\{X_{t,i}\}_{\alpha_i=1}; C_t\right).$$

A scene-level depth test then produces

$$M_t = \text{Vis}\left(\widetilde{M}_t; \{X_{t,i}\}_{i=1}^{N}, C_t\right).$$

Thus, patches occluded by other scene geometry are absent from M_t.

Adapter injection. The camera-conditioning module encodes the camera trajectory C using a relative-ray representation. This encoding is left implicit in the notation $\phi_C^{(\ell)}(C)$ and in the model input C. At transformer block ℓ, the conditioning features are added as

$$h^{(\ell)} \leftarrow h^{(\ell)} + \phi_C^{(\ell)}(C) + \phi_W^{(\ell)}(W) + \phi_M^{(\ell)}(M).$$

The M term is omitted during the first training stage.

Initialization from multiple observations. Although a single frame is sufficient to initialize the initial state, patches from additional observations can be lifted into the same world coordinate system to extend its spatial coverage. The long sequence in Fig. 3(b) uses an initial state initialized from multiple observations and spans three successive rollouts with object control and camera control.

A.2 SAJG CONSTRUCTION

Condition rendering. For each target camera C_t, the scene-anchor condition renders the initial state without interventions:

$$(W_{s,t}, _) = \mathcal{R}(S_0; C_t). \tag{4}$$

The joint condition renders the intervened state:

$$(W_{j,t}, M_t) = \mathcal{R}(S_t; C_t). \tag{5}$$

Both conditions use the complete scene for depth testing. The joint branch receives the intervention-mask condition, while the scene-anchor branch omits it.

Scene-mask construction. We construct G_s in latent-video coordinates. For each target camera, we project the selected entities at their source and target poses to obtain the binary support volumes E_{src} and E_{tgt}. The scene mask is

$$G_s = \mathcal{F}_{1,1}\left[1 - \mathcal{D}_1\left(E_{\text{src}} \cup E_{\text{tgt}}\right)\right], \tag{6}$$

where $\mathcal{D}_1$ applies one-cell spatial dilation and $\mathcal{F}_{1,1}$ applies one-cell spatial and temporal feathering. The resulting values lie in $[0, 1]$, and the mask is broadcast across latent channels. Unlike the intervention mask M, which marks the visible target entity regions, G_s selects scene regions away from both source and target entity supports, with soft transitions at their boundaries.

The mask is geometry-derived rather than learned. It may vary across video frames, but is computed only once for each input and remains unchanged across denoising steps. All dilation and feathering settings are fixed; only α is varied in the routing ablation.

Equivalent correction form. Substituting Eq. (2) into Eq. (3) gives

$$\begin{aligned} V_{\text{final}} &= V_j + (1-\alpha) G_s \odot (V_s - V_j) \\ &= w_s \odot V_s + w_j \odot V_j, \end{aligned} \tag{7}$$

where

$$w_s = (1-\alpha) G_s, \qquad w_j = 1 - (1-\alpha) G_s. \tag{8}$$

Thus, $w_s + w_j = 1$ and $w_j \geq \alpha$ at every spatio-temporal location. The equivalent correction form shows that the scene-anchor prediction modifies the joint prediction only where the scene mask permits. At the default $\alpha = 0.25$, the effective weights are $w_s = 0.75 G_s$ and $w_j = 1 - 0.75 G_s$. Setting $\alpha = 0$ applies the maximum scene correction permitted by the mask, while $\alpha = 1$ removes the correction.

Classifier-free guidance and evaluation cost. Given the same noisy latent x_τ at flow-matching time τ and camera condition C, the conditional video model f_θ produces velocity predictions with classifier-free guidance (CFG):

$$V_s = \text{CFG}\big[f_\theta(x_\tau, \tau; C, W_s)\big], V_j = \text{CFG}\big[f_\theta(x_\tau, \tau; C, W_j)\big]. \tag{9}$$

Both predictions share the source observation and text prompt; these and the intervention-mask input are omitted from the notation.

The main-paper equations use predictions after classifier-free guidance. In implementation, we apply the same spatial weights w_s and w_j separately to the conditional and null-text predictions, and then apply classifier-free guidance to the two mixed predictions. With the same guidance scale for both branches, this is algebraically equivalent to mixing the guided predictions because both operations are linear. The exact computation uses a separate null-text prediction for each branch and requires four transformer evaluations per denoising step.

We additionally tested an approximation that reuses a single null-text prediction for both branches while retaining their separate conditional predictions. Although the two branch-specific null-text predictions are not necessarily identical, we observed no noticeable differences in the generated results under this approximation in the tested settings. Reusing one null-text prediction reduces the computation to three transformer evaluations per denoising step.

A.3 TRAINING

Training settings and captions. Both stages use AdamW with a global batch size of 8 and a base learning rate of 1×10^{-4}. We generate training-video captions using Qwen2.5-VL (Bai et al., 2025). The captions describe scene content while omitting motion and object-location descriptions (e.g., "A is to the left of B").

Optimization. Both stages use the flow-matching objective

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{x,\tau}\left[\left\| f_\theta(x_\tau, \tau; C, W, M) - V_\tau^{\mathrm{gt}} \right\|_2^2\right], \tag{10}$$

with the M condition omitted in the first stage. The first stage trains the C and W conditioning modules on DL3DV for 20K steps. The second stage initializes these modules from the first-stage checkpoint, adds the M module, and jointly trains all three conditioning paths on HM-World for 10K steps. The video backbone, VAE, and text encoder remain frozen.

Training condition construction. We render S_0 from the target camera C_t and zero out the selected entities' projected source regions, obtaining W_t^{src}. Let z_t^e denote the ground-truth entity latent aligned with the target latent grid, and B_t its binary spatial support. We form the training condition as

$$W_t^{\mathrm{train}} = (1 - B_t) \odot W_t^{\mathrm{src}} + B_t \odot z_t^e, \tag{11}$$

with B_t broadcast across latent channels. This retains source-scene appearance outside the target entities, while vacated source regions remain zero unless covered by a target entity.

Conditioning augmentation. For a latent grid of size $H \times W$, $\mathcal{D}_{\mathrm{NN}}$ reduces the spatial resolution to $\mathrm{round}(\rho H) \times \mathrm{round}(\rho W)$ using nearest-neighbor sampling. $\mathcal{U}_{\mathrm{NN}}$ restores the original resolution using nearest-neighbor upsampling. Given the ground-truth entity latents z_t^e and their binary spatial support B_t, we define

$$\widetilde{z}_t^e = \begin{cases} z_0^e, & t = 0, \\ (1 - B_t) \odot z_t^e + B_t \odot \mathcal{U}_{\mathrm{NN}}\big(\mathcal{D}_{\mathrm{NN}}(z_t^e)\big), & t > 0. \end{cases} \tag{12}$$

As defined in Eq. (12), the resampled values replace the conditioning appearance only within the future-frame entity support B_t. The support mask itself is not resampled. The augmented entity latents are then composited with the initial-state renders using Eq. (11). Neither the intervention mask M nor the ground-truth training target is modified.

A.4 WCPE EVALUATION

Motivation and target projections. An object's image-space position alone does not establish whether the prescribed world-space control is realized relative to the scene. WCPE therefore evaluates object–background relative positions together with background alignment. Target trajectories are constructed from the input scene proxy: object points follow the prescribed object transformations, while background points remain fixed in world coordinates. Both are projected using the target camera trajectory. These targets are shared across methods for each evaluation condition.

Metric formulation. We track object queries and background anchors in the generated videos using CoTracker3 (Karaev et al., 2025). For an object query at a given frame, let $\hat{p}$ and p^* denote its tracked image position and target projection. Similarly, $\hat{b}_j$ and b_j^* denote the tracked and target positions of background anchor j. Using the same valid anchor set $\mathcal{V}$, we define

$$e_{\mathrm{rel}} = \mathrm{median}_{j \in \mathcal{V}} \frac{\left\| (\hat{p} - \hat{b}_j) - (p^* - b_j^*) \right\|_2}{D}, \tag{13}$$

$$e_{\mathrm{bg}} = \mathrm{median}_{j \in \mathcal{V}} \frac{\left\| \hat{b}_j - b_j^* \right\|_2}{D}, \tag{14}$$

where $D = \sqrt{W^2 + H^2}$ is the image diagonal. The joint error is

$$e_{\mathrm{joint}} = \max\left(e_{\mathrm{rel}}, e_{\mathrm{bg}}\right). \tag{15}$$

A low joint error requires both accurate object–background relative positions and accurate background alignment with the target projection.

Table 4: Staged versus interleaved training under matched data exposure and computation.

Training configuration	WCPE (%) ↓	Cam-T ↓	Cam-R (°) ↓
Interleaved training	2.630	0.0117	0.413
Staged curriculum (Ours)	**2.323**	**0.0106**	**0.376**

Failure handling and aggregation. We report WCPE@5 with an error cap and missing-measurement penalty of $\tau = 0.05$, corresponding to 5% of the image diagonal:

$$e_{\text{score}} = \begin{cases} \min(e_{\text{joint}}, \tau), & \text{if the measurement is valid,} \\ \tau, & \text{otherwise.} \end{cases} \tag{16}$$

The eligible query–frame pairs are determined by target visibility and finite target projections, independently of the generated video. Invalid measurements remain in this fixed evaluation population and receive the maximum penalty.

Each video-level score is the mean of e_{score} over its eligible query–frame pairs. We evaluate 277 scenarios with two independent random seeds per scenario, yielding 554 videos per method. The final WCPE is the equally weighted mean of all 554 video-level scores, reported in percentage units. The same evaluation protocol is applied to all methods.

Interpretation. WCPE measures projection-level consistency with the prescribed world-space object and camera controls. It does not reconstruct or directly verify metric 3D trajectories, and its accuracy depends on the scene proxy and point tracking.

B Additional ablations

B.1 Curriculum training

We compare the staged curriculum in Section 3.3 with an interleaved baseline that mixes the same conditioning tasks throughout training. Both schedules use 30K total updates: 20K on DL3DV with $C{+}W$ conditioning and 10K on HM-World with $C{+}W{+}M$ conditioning. All other training settings are identical, and we select the best checkpoint for each schedule based on validation performance.

As shown in Table 4, staged training achieves lower WCPE, Cam-T, and Cam-R. Since both schedules use the same data exposure and update counts, this comparison supports the benefit of training order. We also continued the interleaved baseline beyond 30K steps and observed no further validation improvement.

C More comparisons

In this section, we provide detailed representation, capability and qualitative comparison results. The green boxes and XYZ axes indicate the target object positions and orientations, respectively. See `supplementary.html` for the full sequence.



Figure 6: **Qualitative comparison of object translation and rotation control**. The user translates the yellow robot forward and left while rotating it in yaw, and independently translates the camera to the right. Ours follows the robot's approach and turn under the prescribed viewpoint change. VerseCrafter shows insufficient forward motion; MotionCanvas and SymphoMotion deviate from the target heading; and WorldDirector and Real2SAM2Real produce excessive enlargement accompanied by unintended scene motion. See `supplementary.html` for the full sequence.

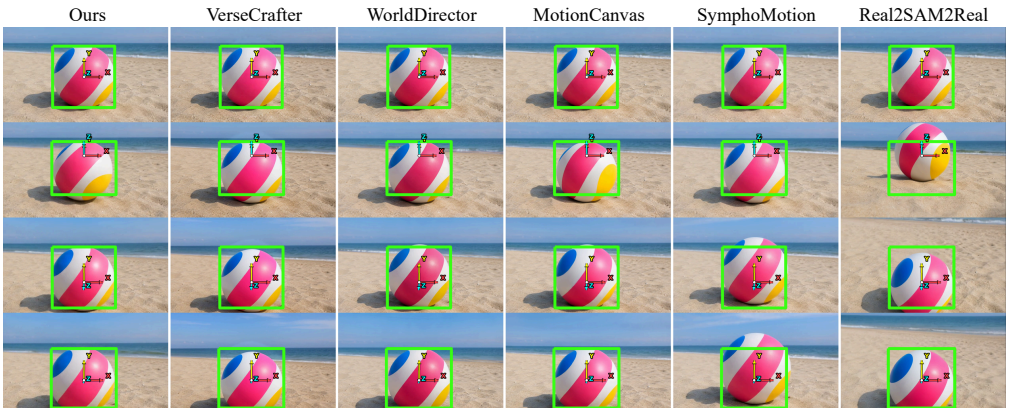


Figure 7: **Qualitative comparison of object rotation control**. The user specifies vertical oscillation and pitch rotation for the ball while pitching the camera upward. VerseCrafter, WorldDirector, SymphoMotion, and Real2SAM2Real fail to reflect the prescribed ball rotation. Ours reproduces the ball's changing orientation and vertical motion under the moving viewpoint. See `supplementary.html` for the full sequence.

Table 5: Capability comparison of object-control methods. Appearance-bearing scene memory denotes spatially grounded appearance evidence that can be reused across viewpoints. Explicit 3D object control indicates that object position or pose can be directly specified in 3D. Occlusion-aware appearance transport requires transporting appearance evidence with explicit depth and visibility reasoning. ✔, ▲, and ✘denote full, partial or indirect, and unsupported capabilities, respectively.

Method	Appearance-bearing scene memory	Explicit 3D object control	Occlusion-aware appearance transport
VerseCrafter	✔	✔	▲
WorldDirector	✘	▲	✘
MotionCanvas	✘	▲	✘
SymphoMotion	▲	✔	▲
Real2SAM2Real	✘	✔	✘
Ours	✔	✔	✔



Figure 8: **Qualitative comparison of occlusion handling**. The user moves the tractor forward behind the hay bales and then partially back while dollying the camera forward. WorldDirector, MotionCanvas, and SymphoMotion cause the hay bales to disappear. Real2SAM2Real absorbs the tractor's intended motion into unintended viewpoint change, while VerseCrafter fails to reproduce its partial return. Ours follows both the tractor trajectory and camera dolly while preserving the bales and their occlusion of the tractor. See `supplementary.html` for the full sequence.

Table 6: Detailed comparison of representations used for user intervention. Geometry denotes the representation used to instantiate the intervention, while appearance indicates whether the representation explicitly carries appearance information. A dash indicates no explicit appearance information.

Method	User intervention	
	Geometry	Appearance
VerseCrafter	3D Gaussian trajectories	–
WorldDirector	2D box trajectories from 3D plans	Box-aligned RGB crops
MotionCanvas	2D boxes / point tracks	–
SymphoMotion	3D point trajectories	–
Real2SAM2Real	3D mesh cache $\rightarrow$ normal maps	–
Ours	**Transformed 3D patches**	**Latent appearance grids**

Table 7: Comparison with MosaicMem under the same evaluation conditions as UniState. MosaicMem supports camera control but does not natively support object control. We apply time-varying 6-DoF transformations to selected object patches for comparison. Under these interventions, we observe diverse failure modes, including camera-control failures and objects remaining stationary or disappearing instead of following the specified motion, as illustrated in Fig. 9.

Method	WCPE (%) $\downarrow$	Cam-T $\downarrow$	Cam-R (°) $\downarrow$
MosaicMem (reproduced)	2.522	0.01249	0.5132
Ours	**2.3233**	**0.0106**	**0.3761**

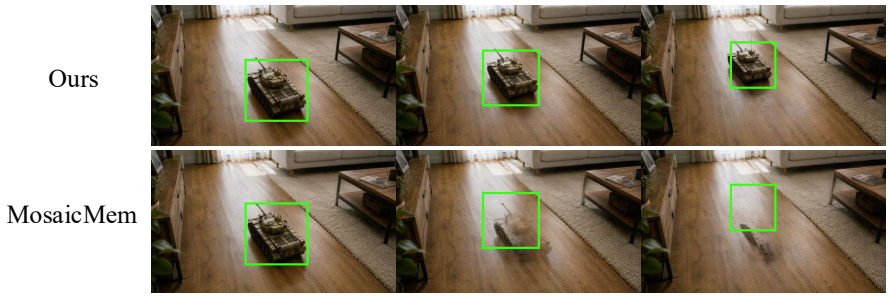


Figure 9: Qualitative failure of MosaicMem in an object-only intervention scenario with a fixed camera. The tank fails to follow the specified motion and disappears midway through the generated sequence.



Figure 10: Misalignment in source-transport conditions. Rows show the source-transport condition (top) and ground-truth video (bottom) for the same sequence. Source transport produces floating artifacts due to inaccurate geometric alignment. TAA instead augments ground-truth entity latents to introduce transport-like appearance distortions while preserving the target entity support, avoiding the misaligned floaters present in the source-transport condition.