MOMENTUM: PERSISTENT DYNAMICS FOR VIDEO WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video world models must capture how a scene evolves beyond the current camera view to generate consistent observations when its content reappears. Existing approaches rely on observation history, static point clouds, or image caches constructed from past frames. These representations are often expensive and, more importantly, cannot update dynamic scene content outside the field of view: generated vehicles and pedestrians may be forgotten after leaving the field of view or reappear in physically inaccurate states. We represent an evolving scene through synchronized videos from reference-camera rigs, each comprising eight cameras that jointly cover the surrounding scene. We connect these rigs in a graph and propagate video generation between neighboring rigs to coordinate scene dynamics across locations and time. We introduce Momentum, a two-stage pipeline: WorldUpdate jointly generates reference-camera videos as a shared memory of scene dynamics, and WorldInteract grounds camera-controlled video generation in this evolving memory, allowing moving content to remain represented outside the target view and reappear consistently. To support training and evaluation, we develop a CARLA pipeline that captures synchronized videos from reference camera rigs and a moving target camera, together with an evaluation framework measuring consistency within the generated world state and the target observations. Experiments on generated street scenes show that grounding target videos in scene dynamics generated by Momentum improves dynamic scene consistency by 126.2% in unseen scenes over static image conditioning, while also improving visual fidelity. Our models and source code will be made publicly available.

1 INTRODUCTION

Video world models simulate how visual environments evolve over time and respond to actions, supporting applications in game generation (Team et al., 2026; Yu et al., 2025b; Qian et al., 2026), robotic learning and control (Gao et al., 2026a; Ye et al., 2026), autonomous driving simulation (Wayve, 2026; Basant et al., 2026), and virtual-world exploration (Ball et al., 2025; Team et al., 2026; Sun et al., 2025). In these settings, the generated video reveals only a partial view of an environment whose contents and dynamics extend beyond what is currently visible. As the viewpoint changes, the generated observations should remain consistent with both the scene's spatial structure and the motion of objects within it. A vehicle approaching an intersection should follow the same course whether the generated video follows it, shows it from another viewpoint, or turns away and later shows it again. This requires grounding generation to a world representation that evolves through time.

Existing memory mechanisms provide part of this grounding through observation history or persistent spatial representations. Conditioning on past frames preserves local continuity, but a finite context limits how much of the environment remains accessible (Zhang et al., 2026; Savov et al., 2026; Cai et al., 2026). Image caches (Xiao et al., 2026; Chen et al., 2026b; Yu et al., 2025a), depth maps (Ren et al., 2025; Shen et al., 2026), and reconstructed point clouds (Ren et al., 2025; Li et al., 2025c) extend spatial memory by retaining information from previously observed locations. However, stored observations describe the scene at the time they were captured. Without a mechanism for advancing their contents, they cannot provide temporally updated grounding for a moving camera.

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

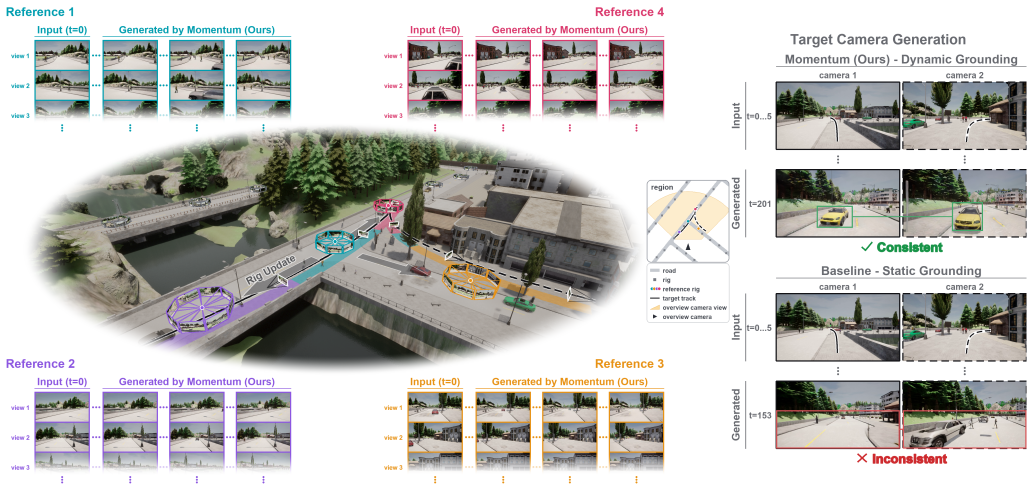

Figure 1: **Momentum generates temporally and cross-view consistent target cameras from sparse camera rig references.** An intersection is observed by four 8-camera reference rigs (*References 1–4*, colour-coded in the scene and shown over time in the corner panels); grey rigs mark further panoramas in the region that this example does not condition on. Two moving target cameras (black line tracks) traverse the junction, each grounded at every step by its nearest reference, indicated by the colour of the band on the road beneath it. Right: camera views generated over time for our method and a baseline, from the same t=0 ground-truth frame. Inset: plan view of the region with the footprint of the overview camera.

The challenge is therefore to maintain a representation that specifies both what exists at a location and how its state changes over time.

Recent approaches have begun to model dynamics beyond the current camera view through explicit object-level motion plans or entity-centered simulation within a static background (Wang et al., 2026a; Duan et al., 2026). These advances support persistent dynamics for selected entities, while maintaining the evolution of general scene content remains relatively unexplored. We pursue a representation that evolves scene content across a coordinated network of reference cameras, without requiring explicit object trajectories, identity tracking, or a static-dynamic scene decomposition.

Seoul World Model (SWM) offers a useful foundation through image-based spatial grounding (Seo et al., 2026). It conditions camera-controlled video generation on retrieved street-view images, using geometric reprojection and semantic referencing to anchor generated observations to a particular environment. These references ground the scene's spatial structure and appearance, while motion is synthesized within the generated video. We extend this principle to *spatiotemporal grounding*: synchronized videos from spatially distributed reference cameras provide a shared account of both scene appearance and motion. The target camera is thus grounded in the scene at the corresponding location and time, allowing different camera trajectories to observe the same unfolding dynamics.

We introduce Momentum, a two-stage pipeline in which WorldUpdate generates an evolving world state and WorldInteract uses it to ground camera-controlled video generation. The method is presented on Fig. 1. The state is organized as a graph of panoramic rigs, each comprising eight reference cameras. WorldUpdate jointly generates videos for the eight cameras of each rig and propagates generation through the graph, conditioning each rig on geometrically reprojected video from an already generated neighbor. This couples the evolution of scene content across reference viewpoints. WorldInteract then generates video along a target camera trajectory using the resulting world rollout as spatiotemporal grounding. The reference videos therefore serve both as the simulation of scene dynamics and as their persistent representation, enabling motion outside the target camera's view to remain part of the state used to generate subsequent observations.

To develop this pipeline, we build an SWM-based image-grounded model on Cosmos Predict 2.5 and a CARLA data-generation framework that provides training and evaluation data for both stages. The framework captures static reference images and synchronized videos from fixed reference-camera

rigs to supervise world-state generation. It also records videos along diverse moving target-camera trajectories, providing supervision for camera-controlled generation grounded in the reference observations. These observations provide supervision for generating a dynamic world state and for rendering camera-controlled video consistent with that state. We develop an evaluation framework for dynamic grounding, combining visual, camera-control, and scene-consistency metrics to assess the evolving world representation and the consistency of target videos with its dynamics. We evaluate our SWM-based baseline on real-world driving data and assess dynamic grounding on CARLA scenes from seen and unseen towns. Generating reference videos introduces an additional inference stage, trading computation for a shared representation of scene dynamics beyond the current view. This cost can be amortized across multiple target-camera trajectories grounded in the same world rollout, while improving dynamic consistency over static image grounding. With ground-truth reference videos, the Dynamically Grounded WM improves visual fidelity, camera control, and scene consistency over static grounding across single- and multi-rig settings. End-to-end, Momentum generates its own reference videos and improves dynamic scene consistency by 130.2% in seen towns and 126.2% in an unseen town compared with the Dynamically Grounded WM conditioned on static images.

Our contributions are threefold:

- We introduce a spatiotemporal world representation comprising a graph of reference-camera rigs with synchronized multi-view videos, providing shared spatial and dynamic grounding across camera viewpoints and time.
- We propose Momentum, a two-stage pipeline that generates scene dynamics through graph-propagated multi-view generation in WorldUpdate and grounds camera-controlled observations in those dynamics through WorldInteract.
- We develop a CARLA data-generation pipeline for synchronized reference and target camera observations, an evaluation framework for assessing dynamic scene consistency in both the generated world representation and the resulting target videos, and an image-grounded world model based on SWM.

2 Related Work

2.1 Video World Models

Building on learned models of environment dynamics (Ha & Schmidhuber, 2018), generative video world models increasingly support interactive visual environments across domains. Genie (Bruce et al., 2024; Ball et al., 2025), LingBot-World (Team et al., 2026), and HY-WorldPlay (Sun et al., 2025) enable controllable exploration. DIAMOND (Alonso et al., 2024), GameFactory (Yu et al., 2025b), and Matrix-Game 3.5 (Qian et al., 2026) develop interactive game-like simulation, while Pandora (Xiang et al., 2024) uses natural-language actions to guide video evolution. Driving models such as GAIA (Wayve, 2026) and OmniDreams (Basant et al., 2026) generate controllable observations for simulation and evaluation. In robotics, DreamDojo (Gao et al., 2026a) learns action-conditioned simulation from human-video priors, DreamZero (Ye et al., 2026) jointly predicts future video and robot actions for control, and RoboWorld (Jeon et al., 2026) supports robot-policy evaluation through video simulation. Cosmos (Agarwal et al., 2025; 2026) provides pretrained world-model capabilities for physical AI; we build on Cosmos Predict 2.5 (Ali et al., 2025). Our focus is on organizing generated dynamics into a persistent world state that grounds observations across target-camera trajectories.

2.2 Memory and Dynamics in Video World Models

Memory mechanisms extend generation beyond a short observation window. StateSpaceDiffuser (Savov et al., 2026) uses recurrent state-space modeling, FramePack (Zhang et al., 2026) compresses frame history, and Mixture of Contexts (Cai et al., 2026) learns sparse attention over informative temporal contexts. WorldMem (Xiao et al., 2026), VRAG (Chen et al., 2026b), Context-as-Memory (Yu et al., 2025a), and MemCam (Gao et al., 2026b) retrieve historical observations. Spatial organization further supports consistency through 3D caches in GEN3C (Ren et al., 2025), surfel indexing in VMem (Li et al., 2025c), and geometric memory in SPMem (Wu et al., 2026), Captain Safari

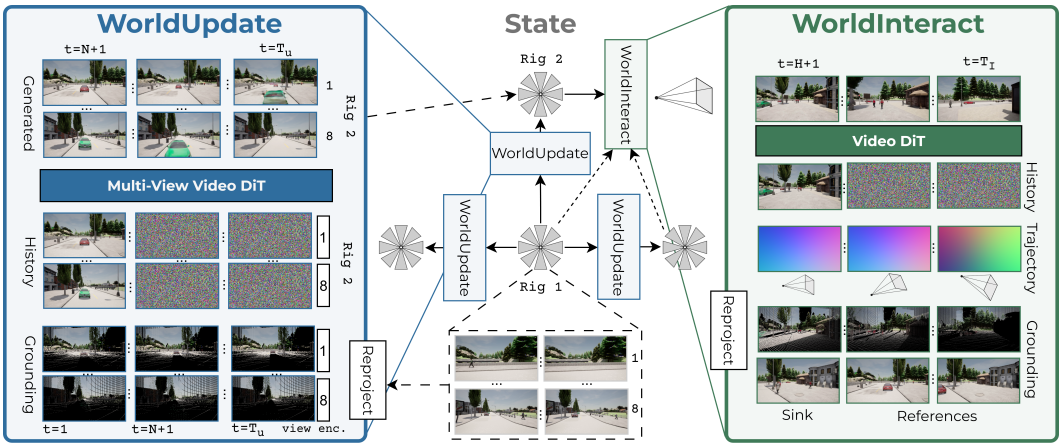

Figure 2: **Architecture of Momentum.** WorldUpdate generates synchronized reference-camera videos as an evolving world state, which WorldInteract uses to ground target-camera generation.

(Chou et al., 2026), and Lyra 2.0 (Shen et al., 2026). Latent Spatial Memory (Wang et al., 2026b) stores spatial information in diffusion latent space, while SWM (Seo et al., 2026) grounds generation in external street-view images. These mechanisms retain past evidence, but memory retention alone does not advance the represented scene through time.

Dynamic geometric methods such as NVFi (Li et al., 2023), FreeGave (Li et al., 2025b), and TRACE (Li et al., 2025a) predict future RGB views by fitting scene dynamics to recorded videos. We instead reuse a dynamics model across scenes and spatial regions without scene-specific fitting. InfiniCube (Lu et al., 2025) and WorldSplat (Zhu et al., 2026) generate dynamic environments under supplied object-motion controls; our model generates scene evolution without prescribed object trajectories.

Persistent dynamics beyond the target view remain relatively unexplored. HyDRA (Chen et al., 2026a) retrieves motion and appearance cues for hidden subjects, while MosaicMem (Yu et al., 2026) and Mem-World (Zheng et al., 2026) combine spatial memory with dynamic video generation. WorldDirector (Wang et al., 2026a) plans object trajectories through an LLM, and LiveWorld (Duan et al., 2026) advances active entities through virtual monitors against a static background. Momentum instead organizes its evolving state by spatial regions: coordinated reference-camera videos generate and represent scene dynamics, then ground target-camera observations. This formulation supports spatial extension through additional rigs without explicit entity tracking or a static–dynamic decomposition.

3 Methodology

3.1 Problem Statement

Given a history of H observations $I_{1:H}$, the corresponding camera poses $c_{1:H}$, future camera trajectory $c_{H+1:H+T}$, and a world grounding G, a grounded camera-controlled video world model $f(I_{1:H}, c_{1:H+T}, G)$ predicts the corresponding future observations $\hat{I}_{H+1:H+T}$, where G provides additional information about the scene beyond that contained in the observation history, such as spatial context or previously unobserved regions of the environment.

Existing grounded world models such as SWM use G as a collection of static spatial references, which may originate from different points in time and therefore do not constitute a temporally evolving representation of the dynamic scene. We instead model the grounding as a time-indexed world state $S_{H:H+T}$, allowing the grounding itself to evolve throughout generation. The prediction can then be written as $\hat{I}_{H+1:H+T} = h(I_{1:H}, c_{H+1:H+T}, S_{H:H+T})$, where the state sequence is obtained from an initial grounded state through $S_{H:H+T} = g(S_H)$.

In our model, S_t is represented by N spatially distributed reference-camera rigs, each comprising eight co-located cameras with uniformly spaced orientations. Thus, $S_t = \{I_t^{n,k}\}_{n=1,\dots,N;\, k=1,\dots,8}$. Unlike the static spatial references used in SWM, all reference cameras are temporally synchronized and generate video, so $S_{H:H+T}$ forms a spatiotemporally evolving representation of the scene.

3.2 Data Generation Framework

Training and evaluating our models requires reference observations paired with target-camera trajectories, with synchronized video for dynamic grounding. No large-scale public corpus meets these requirements, and SWM's real-world data is static and proprietary. We therefore develop an open CARLA (Dosovitskiy et al., 2017) framework for static and dynamic grounding, across six towns with randomized weather, illumination, vehicles, and pedestrians. Further details - in App. A.1.

Static grounding data. We pair target videos with panoramic reference rigs covering the same maps. Each rig contains eight outward-facing cameras spaced 45° apart in yaw, positioned 2 m above the road. Rigs follow the road graph at 20 m intervals with small random positional offsets. Captures are separated by 2.4 s, equivalent to 30 km/h travel, allowing dynamic content to evolve between neighboring captures as in street-view acquisition. Target videos span 36 s at 10 fps along vehicle (autopilot), pedestrian (AI-controlled walker), or flyover (freely moving elevated camera) trajectories. References and targets are captured in independent sessions with independently sampled conditions, sharing static structure while differing in transient content.

Dynamic grounding data. To supervise evolving reference views, we capture reference and target videos in a shared simulation rollout. A moving target camera traverses a randomly selected intersection covered by one panoramic rig (*single-rig*) or up to four spatially distributed rigs (*multi-rig*). Each session spans 300 frames at 10 fps, with synchronized observations depicting the same scene evolution. To reduce storage, rigs record new observations only when the target camera is within 25 m; otherwise, their latest frames are repeated to retain temporal alignment. Active intervals provide eight synchronized streams per rig for world-state supervision, while target trajectories revisit shared content from different viewpoints and headings to supervise grounded camera-controlled generation.

We collect 20 sessions per setting per town for training and two evaluation sets per setting of 30 sessions each: Test 1 uses unseen intersections in training towns, and Test 2 uses a town excluded from training to assess generalization.

3.3 Static Grounding

Our ***Statically Grounded World Model*** follows the SWM architecture (Seo et al., 2026). As it is currently not publicly available, we build an open implementation of it. Details appear in App. A.2.

Backbone and rollout. The backbone is a Diffusion Transformer operating in the latent space of a causal 3D VAE, initialized from Cosmos Predict 2.5-2B (Ali et al., 2025). Generation proceeds autoregressively in chunks, denoising chunk i conditioned on a text prompt $P^{(i)}$ (we use a constant preset prompt), target-camera trajectory $C^{(i)}$, and H history latents from chunk $i-1$. Per-frame Plücker ray maps, encoded by a shallow convolutional encoder and added to latent tokens, provide camera poses for history, target, reference, and sink latents. We train under teacher forcing on the CARLA data of Sec. 3.2.

Retrieval-augmented grounding. Each chunk retrieves up to K reference images near the target trajectory, with camera poses and metric depth supplied by CARLA during training and estimated with Depth Anything V3 (Lin et al., 2025) at inference. Two pathways incorporate these references. *Geometric referencing* unprojects the nearest references for each target frame into a point cloud and forward-splats it into the target viewpoint; the resulting warped video is encoded and concatenated channel-wise with noisy target latents for dense spatial conditioning. *Semantic referencing* encodes each reference into a single latent, appended to the token sequence at a RoPE temporal position separated from the generation window by G_{ref}. Target latents attend to all K references to recover appearance information lost through depth errors or occlusion.

A *virtual lookahead sink* adds a retrieved latent at RoPE position Δ_{VL} beyond the chunk as a clean anchor, refreshed each generation step. References and targets come from different scene realizations, encouraging reliance on persistent structure rather than transient content. Sec. 4 compares this baseline with SWM's reported scores.

3.4 MOMENTUM

Momentum is a two-stage pipeline that generates an evolving world state and uses it to ground camera-controlled video. The architecture of our method is shown on Fig. 2. The state S_t comprises the reference-camera rig views, advanced through time rather than held fixed. WorldUpdate implements the state transition $S_{H:H+T} = g(S_H)$, generating synchronized videos across rigs to represent scene dynamics independently of the target-camera trajectory. WorldInteract implements the observation model h, generating target-camera video conditioned on the observation history, camera trajectory, and corresponding world state. The reference videos thus serve both as the simulation of scene dynamics and as their persistent representation, grounding observations when moving content leaves and later re-enters the target view.

Dynamics are represented without explicit entity detection, identity tracking, or re-identification, so for a fixed rig configuration the generation workload does not grow with the number of moving entities. The two stages operate independently: we store completed world rollouts in the dataset format of Sec. 3.2, allowing WorldInteract to consume generated and recorded reference videos through the same interface. A single world rollout can ground multiple target-camera trajectories, be inspected independently, and be generated at a different resolution or on a different machine from the target video.

3.4.1 WORLDUPDATE

WorldUpdate is the component of Momentum responsible for advancing the world state through synchronized reference-camera video generation. It requires consistency both across cameras within a rig and across spatially distributed rigs observing shared content. To meet these requirements, we introduce the ***Multi-Rig Dynamically Grounded World Model*** (Multi-Rig DGWM) as an architecture for this component, combining joint multi-view generation with geometric conditioning between rigs.

Multi-view generation. Multi-Rig DGWM builds on the pretrained multi-view video diffusion transformer of Cosmos-Predict 2.5 (Ali et al., 2025), trained on autonomous driving data. It extends its seven-view configuration to the eight cameras of our rigs by adding an extra learnable view embedding. View-stacked attention exchanges information across cameras during joint generation, supporting cross-view consistency. We incorporate the warped-video channels of geometric referencing and the Plücker camera residual from the grounded model. Semantic reference tokens, the virtual lookahead sink, and their associated attention masks are omitted: each fixed camera's generated history anchors its viewpoint and appearance, while reprojected neighbor videos provide spatially aligned conditioning on scene evolution.

Multi-rig consistency. Independent rig generation can produce conflicting depictions of shared content, including objects duplicating or disappearing between viewpoints. We therefore condition each rig on a neighbor's synchronized video, unprojecting its views with depth and reprojecting them into the receiving rig using the known relative poses. During training, conditioning comes from the ground-truth video of the nearest other rig, allowing training samples to be processed independently. At inference, neighboring videos must themselves be generated, requiring a schedule that makes each conditioning source available before its dependent rig.

Graph propagation. We construct a graph connecting rigs close within a range to provide useful reprojection conditioning. Generation proceeds in chunks of one model window, overlapping by the conditioning frames to maintain temporal continuity. Within each chunk, we traverse the graph breadth-first from a source rig, which generates from its own history without neighbor conditioning. Each subsequent rig conditions on the freshly generated video of its nearest already-generated neighbor(s), while rigs at equal depth are generated together. This propagates scene evolution across the graph over a shared temporal interval.

Table 1: **Baseline on Ann-Arbor-City-Bench (MARS).** SWM scores are those reported by the authors; our statically grounded model is trained on synthetic data alone. Masked PSNR and LPIPS are computed on static regions only, following the SWM protocol. Best value per metric in bold.

Method	mPSNR ↑	mLPIPS ↓	RotErr ↓	TransErr ↓
SWM Seo et al. (2026) (reported)	15.18	0.481	0.055	**0.154**
Statically Grounded World Model (ours)	**16.30**	**0.426**	**0.025**	0.238

Table 2: **WorldInteract consistency across scene configurations.** Models use ground-truth reference observations from scenes containing one or four rigs. Image conditioning uses fixed initial observations; video conditioning uses synchronized frames. Best scores for each scene configuration are **bold**; best scores across configurations within each test set are underlined.

Method	Cond.	PSNR ↑		LPIPS ↓		RotErr ↓		TransErr ↓		SCS-static ↑		SCS-dyn ↑	
rigs in the scene:		1 rig	4 rigs	1 rig	4 rigs	1 rig	4 rigs	1 rig	4 rigs	1 rig	4 rigs	1 rig	4 rigs
test 1: seen towns													
Statically grounded WM	image	21.01	19.89	0.436	0.426	0.031	0.038	0.181	0.160	0.512	0.526	0.056	0.078
Statically grounded WM	video	20.97	20.35	0.436	0.414	0.029	0.030	0.171	0.156	0.502	0.562	0.083	0.127
Dynamically Grounded WM (ours)	video	**23.55**	**23.16**	**0.352**	**0.335**	**0.021**	**0.022**	**0.054**	**0.056**	**0.630**	**0.652**	**0.202**	**0.265**
test 2: unseen town													
Statically Grounded WM	image	17.94	20.27	0.508	0.449	0.026	0.026	0.176	0.148	0.426	0.573	0.057	0.081
Statically Grounded WM	video	17.93	19.23	0.507	0.456	0.026	0.034	0.196	0.167	0.425	0.570	0.059	0.112
Dynamically Grounded WM (ours)	video	**20.36**	**21.94**	**0.425**	**0.371**	**0.023**	**0.023**	**0.047**	**0.053**	**0.583**	**0.667**	**0.147**	**0.230**

Table 3: **WorldInteract consistency across reprojection source counts.** Within four-rig scenes, geometric conditioning combines ground-truth reference observations from one or three source rigs. Best scores for each source count are **bold**; best scores across source counts within each test set are underlined.

Method	Cond.	PSNR ↑		LPIPS ↓		RotErr ↓		TransErr ↓		SCS-static ↑		SCS-dyn ↑	
warp sources:		1 src	3 src	1 src	3 src	1 src	3 src	1 src	3 src	1 src	3 src	1 src	3 src
test 1: seen towns													
Statically Grounded WM	image	19.89	19.55	0.426	0.429	0.038	0.034	0.160	0.173	0.526	0.503	0.078	0.074
Statically Grounded WM	video	20.35	20.04	0.414	0.419	0.030	0.029	0.156	0.172	0.562	0.540	0.127	0.107
Dynamically Grounded WM (ours)	video	**23.16**	**22.89**	**0.335**	**0.337**	**0.022**	**0.020**	**0.056**	**0.053**	**0.652**	**0.647**	**0.265**	**0.230**
test 2: unseen town													
Statically Grounded WM	image	20.27	20.33	0.449	0.449	0.026	0.026	0.148	0.146	0.573	0.563	0.081	0.079
Statically Grounded WM	video	19.23	20.42	0.456	0.430	0.034	0.027	0.167	0.143	0.570	0.589	0.112	0.131
Dynamically Grounded WM (ours)	video	**21.94**	**23.05**	**0.371**	**0.351**	**0.023**	**0.019**	**0.053**	**0.043**	**0.667**	**0.680**	**0.230**	**0.260**

3.4.2 WorldInteract

WorldInteract is the component of Momentum that generates target-camera video grounded in the evolving world state. It must follow the specified camera trajectory while preserving the scene appearance and motion represented by the reference videos. For this component, we introduce the ***Dynamically Grounded World Model*** (Dynamically Grounded WM), retaining the architecture of the Statically Grounded World Model (Sec. 3.3) and fine-tuning it on synchronized dynamic grounding data (Sec. 3.2).

For each chunk, the model retrieves K nearby rig views as semantic references and reprojects the nearest reference view into each target frame. Unlike static grounding, geometric conditioning uses the reference video frame at the corresponding target time, conveying the scene's evolving appearance and motion. Semantic references are instead sampled at the chunk's start, providing a common temporal anchor across the retrieved viewpoints.

4 Experiments

Experimental Setup. We train Statically Grounded WM, Dynamically Grounded WM and Multi-Rig DGWM on the CARLA training set (Sec. 3.2) on 6 towns, according to the training schedule outlined in App. A.2 and App. A.3. We evaluate them in the roles of components of Momentum. We evaluate on Test 1 and Test 2, as collected in Sec. 3.2. Unless otherwise stated, we use the 4-rig version of each set and choose 2 views of 1 closest rig for reprojection. We evaluate visual fidelity

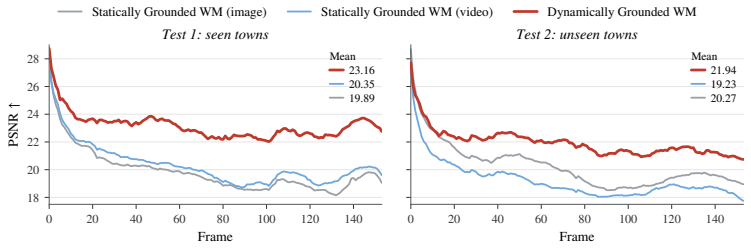

Figure 3: **PSNR per frame over the generation horizon of WorldInteract.**

Table 4: **End-to-end evaluation.** Momentum rolls out the static references in the world state with WorldUpdate and renders the ego view with WorldInteract; Statically and Dynamically Grounded models are directly conditioned on the available static image references.

Method	PSNR ↑	LPIPS ↓	RotErr ↓	TransErr ↓	SCS-static ↑	SCS-dyn ↑
test 1: seen towns						
Statically Grounded WM	19.89	0.426	0.038	0.160	0.526	0.078
Dynamically Grounded WM (image)	22.02	0.358	**0.026**	**0.060**	0.623	0.086
Momentum	**24.03**	**0.341**	**0.026**	0.080	**0.658**	**0.198**
test 2: unseen town						
Statically Grounded WM	20.27	0.449	0.026	0.148	0.573	0.081
Dynamically Grounded WM (image)	22.44	0.380	0.021	0.048	0.658	0.084
Momentum	**22.98**	**0.358**	**0.018**	**0.040**	**0.698**	**0.190**

- PSNR, LPIPS, camera trajectory accuracy - rotation and translation error and scene consistency - SCS-static (Bagchi et al. (2026)), which we adapt to dynamic objects in SCS-dyn. See App. B.

Baseline Validation. We evaluate our Statically Grounded World Model on the publicly available Ann Arbor driving data from MARS (Li et al., 2024), following SWM's evaluation protocol, using masked versions of PSNR and LPIPS (see App. B). This comparison assesses whether our reimplementation provides a competitive baseline despite training exclusively on synthetic data, whereas SWM additionally uses video interpolated from proprietary real-world panoramas. As shown in Tab. 1, our model achieves higher mPSNR, lower mLPIPS, and lower rotation error than the reported SWM results, although translation error remains higher. These results establish a competitive image-grounded baseline for evaluating the benefits of dynamic grounding.

WorldInteract Consistency Evaluation. We evaluate model choices for WorldInteract using ground-truth reference observations, isolating target-video generation from errors in WorldUpdate. We compare the Statically Grounded World Model conditioned on either fixed reference images from $t = 0$ or synchronized reference video against the Dynamically Grounded WM conditioned on video. Image conditioning reprojects the initial observations into each target viewpoint, while video conditioning uses time-aligned observations. Across scenes containing one or four reference-camera rigs, the Dynamically Grounded WM achieves the best visual fidelity, camera-trajectory accuracy, and static and dynamic scene consistency in both seen and unseen towns (Tab. 2). Supplying video conditioning to the statically trained model alone does not yield comparable gains, demonstrating the importance of training with synchronized dynamic grounding.

WorldInteract Reprojection Sources. We next vary the number of rigs supplying conditioning within four-rig scenes, using either one or three source rigs with the same model and conditioning choices. From each source, we select the two camera views closest in orientation to the target camera and combine their reprojections into a single conditioning image per target frame. The Dynamically Grounded WM outperforms both baselines with either source count (Tab. 3). In seen towns, one source provides better visual and semantic consistency, while three slightly improve camera-trajectory accuracy. In the unseen town, three sources improve all metrics, despite training with a single source rig, showing that the model can exploit additional reference coverage at inference.

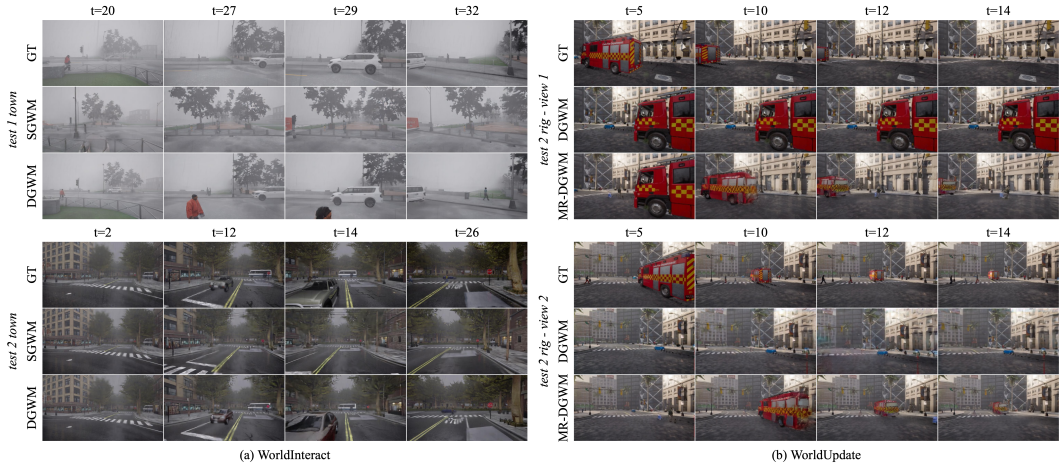

Figure 4: **Qualitative examples.** a) WorldInteract, SGWM versus DGWM in the scenes of both testsets; b) WorldUpdate, DGWM vs Multi-Rig DGWM in two views of the same rig in the test 2.

WorldInteract Performance Over Generation Horizon. We evaluate how visual fidelity changes throughout generation using per-frame PSNR on both test sets. As shown in Fig. 3, the Dynamically Grounded WM maintains visual fidelity more consistently over time than models using static grounding.

End-to-End Evaluation. We evaluate Momentum from static reference images, without ground-truth reference videos. The baselines directly condition target-video generation on these images, while Momentum uses WorldUpdate to generate an evolving world state through graph propagation and WorldInteract to ground target observations in that state. As shown in Tab. 4, Momentum improves visual fidelity and both static and dynamic scene consistency over both baselines in seen and unseen towns. It also achieves the best camera-trajectory accuracy in the unseen town, while matching rotation accuracy but increasing translation error relative to the image-conditioned Dynamically Grounded WM in seen towns. Most notably, SCS-dyn improves over this baseline by 130.2% and 126.2% in seen and unseen towns, respectively, demonstrating the benefit of generating scene dynamics beyond those represented by static references.

Qualitative Results. We qualitatively show examples of WorldUpdate and WorldInteract, comparing static and dynamic grounding, single-view and multi-rig models. In Fig. 4a, it is apparent that the dynamically grounded model preserves moving objects (cars, pedestrians) significantly better across the inference window. In Fig. 4b, the multi-rig model renders objects consistently across views and predicts their movement, while the single-view model is inconsistent and tends to predict no movement (both models are seeded with single frame at t=0).

5 Conclusion

We introduced Momentum, a two-stage pipeline that grounds camera-controlled video generation in an evolving world state. Synchronized reference-camera videos represent scene appearance and dynamics beyond the target view, without explicit entity tracking or a static–dynamic decomposition. WorldUpdate advances this state through graph-propagated generation, while WorldInteract uses it to ground target-camera observations. We developed the Dynamically Grounded WM and Multi-Rig DGWM for these roles, together with a CARLA data-generation framework, an SWM-based Statically Grounded World Model, and an evaluation framework. Experiments in seen and unseen towns demonstrate improved visual fidelity, controllability and scene consistency through dynamic grounding. Compared with the Dynamically Grounded WM conditioned on static image references, Momentum improves dynamic scene consistency (SCS-dyn) by 130.2% in seen towns and 126.2% in the unseen town. These results open a path toward spatially extensible video world models in which multiple viewpoints and interactions share a consistently evolving scene.

6 AI USE STATEMENT

In this work, we used generative AI tools for providing feedback on research methodology or experiments and implementing methods. We have not used generative AI tools for designing research methodology, generating synthetic datasets, helping develop theoretical models or conceptual frameworks, supporting qualitative and thematic data analysis, interpreting results, proposing or refining hypotheses. We have not used an AI tool in any form to formulate the ideas and design our system. Formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, assisting with translation, cleaning and reformatting dataset are not applicable to this work. Additionally, we used generative AI tools for creating or modifying scientific figures or images, creating software code, drafting parts of the research paper and improving its readability. We have reviewed all AI-assisted work, including code and text. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *Advances in Neural Information Processing Systems*, 37:58757–58791, 2024.

Anurag Bagchi, Zhipeng Bao, Homanga Bharadhwaj, Yu-Xiong Wang, Pavel Tokmakov, and Martial Hebert. Walk through paintings: Egocentric world models from internet priors. In *European Conference on Computer Vision (ECCV)*, 2026.

Philip J Ball, Jakob Bauer, Frank Belletti, B Brownfield, A Ephrat, S Fruchter, A Gupta, K Holsheimer, A Holynski, J Hron, et al. Genie 3: A new frontier for world models. *Google DeepMind Blog*, pp. 253–279, 2025.

Aarti Basant, Amlan Kar, Despoina Paschalidou, Fangyin Wei, Francesco Ferroni, Guillermo Garcia Cobo, Haithem Turki, Huan Ling, Jaewoo Seo, James Lucas, et al. Nvidia omnidreams: Real-time generative world model for closed-loop autonomous vehicle simulation. *arXiv preprint arXiv:2606.03159*, 2026.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first international conference on machine learning*, 2024.

Shengqu Cai, Ceyuan Yang, Lvmin Zhang, Yuwei Guo, Junfei Xiao, Ziyan Yang, Yinghao Xu, Zhenheng Yang, Alan Yuille, Leonidas Guibas, et al. Mixture of contexts for long video generation. In *International Conference on Learning Representations*, volume 2026, pp. 92075–92093, 2026.

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

Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

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. *arXiv preprint arXiv:2603.25716*, 2026a.

Taiye Chen, Xun Hu, Zihan Ding, and Chi Jin. Learning world models for interactive video generation. *Advances in Neural Information Processing Systems*, 38:154456–154483, 2026b.

Yu-Cheng Chou, Xingrui Wang, Yitong Li, Jiahao Wang, Hanting Liu, Cihang Xie, Alan Yuille, and Junfei Xiao. Captain safari: A world engine with pose-aligned 3d memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25347–25357, 2026.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pp. 1–16. PMLR, 2017.

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. *arXiv preprint arXiv:2603.07145*, 2026.

Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, et al. Dreamdojo: A generalist robot world model from large-scale human videos. *arXiv preprint arXiv:2602.06949*, 2026a.

Xinhang Gao, Junlin Guan, Shuhan Luo, Wenzhuo Li, Guanghuan Tan, and Jiacheng Wang. Memcam: Memory-augmented camera control for consistent video generation. *arXiv preprint arXiv:2603.26193*, 2026b.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440, 2018.

Byeongguk Jeon, Seonghyeon Ye, JaeHyeok Doo, Sungdong Kim, Minjoon Seo, Hyungmok Son, and Kimin Lee. Roboworld: Fast and reliable neural simulators for generalist robot policy evaluation. *arXiv preprint arXiv:2607.01060*, 2026.

Jinxi Li, Ziyang Song, and Bo Yang. Nvfi: Neural velocity fields for 3d physics learning from dynamic videos. *Advances in Neural Information Processing Systems*, 36:34723–34751, 2023.

Jinxi Li, Ziyang Song, and Bo Yang. Trace: Learning 3d gaussian physical dynamics from multi-view videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8820–8829. IEEE, 2025a.

Jinxi Li, Ziyang Song, Siyuan Zhou, and Bo Yang. Freegave: 3d physics learning from dynamic videos by gaussian velocity. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12433–12443. IEEE, 2025b.

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

Yiming Li, Zhiheng Li, Nuo Chen, Moonjun Gong, Zonglin Lyu, Zehong Wang, Peili Jiang, and Chen Feng. Multiagent multitraversal multimodal self-driving: Open mars dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22041–22051, June 2024.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Yifan Lu, Xuanchi Ren, Jiawei Yang, Tianchang Shen, Zhangjie Wu, Jun Gao, Yue Wang, Siheng Chen, Mike Chen, Sanja Fidler, et al. Infinicube: Unbounded and controllable dynamic 3d driving scene generation with world-guided video models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27272–27283. IEEE, 2025.

Runjia Qian, Zile Wang, Jihai Zhang, Kai Zou, Wei Yu, Jiaxing Li, Zexiang Liu, Yaokun Li, Fei Kang, Kaichen Huang, et al. Matrix-game 3.5: Enhancing real-time streaming interactive world models with patch memory. *arXiv preprint arXiv:2608.29910*, 2026.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6121–6132. IEEE, 2025.

Nedko Savov, Naser Kazemi, Deheng Zhang, Danda Pani Paudel, Xi Wang, and Luc V Gool. Statespacediffuser: Bringing long context to diffusion world models. *Advances in Neural Information Processing Systems*, 38:68865–68898, 2026.

Junyoung Seo, Hyunwook Choi, Minkyung Kwon, Jinhyeok Choi, Siyoon Jin, Gayoung Lee, Junho Kim, JoungBin Lee, Geonmo Gu, Dongyoon Han, et al. Seoul world model: Grounding world simulation models in a real-world metropolis. In *European Conference on Computer Vision*, pp. 562–580. Springer, 2026.

Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, Zian Wang, Nicholas Sharp, Zan Gojcic, et al. Lyra 2.0: Explorable generative 3d worlds. *arXiv preprint arXiv:2604.13036*, 2026.

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. *arXiv preprint arXiv:2512.14614*, 2025.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

Hanlin Wang, Hao Ouyang, Qiuyu Wang, Wen Wang, Qingyan Bai, Ka Leong Cheng, Yue Yu, Yixuan Li, Yihao Meng, Zichen Liu, et al. Worlddirector: Building controllable world simulators with persistent dynamic memory. *arXiv preprint arXiv:2607.02517*, 2026a.

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. *arXiv preprint arXiv:2606.09828*, 2026b.

Wayve. GAIA-4: Multimodal world models powering closed-loop simulation for safe and scalable autonomy. `https://wayve.ai/thinking/gaia-4/`, August 2026. Published August 3, 2026. Accessed September 25, 2026.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *Advances in Neural Information Processing Systems*, 38:49371–49393, 2026.

Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, et al. Pandora: Towards general world model with natural language actions and video states. *arXiv preprint arXiv:2406.09455*, 2024.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *Advances in Neural Information Processing Systems*, 38:49632–49652, 2026.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

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*, pp. 1–11, 2025a.

Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactorly: Creating new games with generative interactive videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11590–11599. IEEE, 2025b.

Wei Yu, Runjia Qian, Yumeng Li, Liquan Wang, Songheng Yin, Dennis Anthony, Yang Ye, Yidi Li, Weiwei Wan, Animesh Garg, et al. Mosaicmem: Hybrid spatial memory for controllable video world models. *arXiv preprint arXiv:2603.17117*, 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. *Advances in Neural Information Processing Systems*, 38:30546–30566, 2026.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Zirui Zheng, Jiaqian Yu, Xiongfeng Peng, Mingyi Li, Chao Zhang, Weiming Li, Dong Wang, Huchuan Lu, Xu Jia, et al. Mem-world: Memory-augmented action-conditioned world models for persistent robot manipulation. *arXiv preprint arXiv:2606.18960*, 2026.

Ziyue Zhu, Zhanqian Wu, Zhenxin Zhu, Lijun Zhou, Haiyang Sun, Bing Wang, Kun Ma, Guang Chen, Hangjun Ye, Jin Xie, et al. Worldsplat: Gaussian-centric feed-forward 4d scene generation for autonomous driving. In *International Conference on Learning Representations*, volume 2026, pp. 68258–68277, 2026.

A Appendix

A.1 Data Generation Implementation Details

All observations are rendered using 1280×720 pinhole cameras with a 90° field of view. Each capture stores RGB video together with per-frame camera extrinsics, intrinsics, and GNSS measurements; depth is stored losslessly. Data collection is configured declaratively, supports resuming interrupted sessions, and can be distributed across multiple CARLA servers.

Static grounding. Panoramic rigs are placed at road waypoints with a random positional offset of up to $1.5\,\mathrm{m}$. Each rig contains eight co-located cameras with yaw offsets spaced by 45° and records both RGB and depth. Target vehicle trajectories use an autopilot vehicle travelling at a randomized fraction below the local speed limit. Pedestrian trajectories use an eye-height camera at $1.65\,\mathrm{m}$ mounted to an AI walker with randomly sampled navigation goals. Flyover trajectories follow collision-checked random walks at 2–8 m altitude, with camera pitch sampled around -20° with $\pm 15^\circ$ variation and additional yaw perturbations. We render 20 sessions per town and trajectory type.

Dynamic grounding. Each session is centred on a randomly selected intersection with at least three approach directions, bounding-box half-extent of at least $5\,\mathrm{m}$, and sufficient nearby vehicle spawn points and pedestrian navigation area within $60\,\mathrm{m}$. In the multi-rig setting, up to four rigs are placed on road waypoints within $40\,\mathrm{m}$ of the intersection and at least $12\,\mathrm{m}$ apart, with one rig located at the junction centre. The single-rig setting retains only the central rig.

We spawn up to 10 vehicles and 25 pedestrians around the selected region. The target camera is placed at an eye height of 1.8–2.2 m and follows a smoothed, collision-checked trajectory visiting the rigs in randomized order. The path remains within $30\,\mathrm{m}$ of the nearest reference rig in the multi-rig setting and $20\,\mathrm{m}$ in the single-rig setting, promoting repeated observation of shared geometry from different headings, including U-turns.

Each session consists of a single synchronous CARLA rollout of 300 frames at $10\,\mathrm{fps}$. Every reference rig records eight RGB and eight depth streams aligned with the target-camera timeline. To reduce rendering cost, reference rigs are active only while the target camera is within $25\,\mathrm{m}$; outside this range, the most recent frame is repeated to preserve temporal alignment. Test sessions use disjoint random seeds, and Test 1 additionally excludes intersections used during training.

A.2 Statically Grounded World Model Implementation Details

Architecture. The backbone is the 2B Cosmos-Predict2.5 DiT: 28 blocks of width 2048 with 16 heads, spatial patch size 2, AdaLN-LoRA of rank 256, and 3D RoPE. The VAE compresses by $4\times$ temporally and $8\times$ spatially into 16 latent channels. Three modifications are made. First, the input embedder is widened from 16 to 32 latent channels so that the warped video of the geometric pathway can be concatenated to the noisy target; with the condition and padding masks, the network sees 34 channels. Second, separate patch embedders, weight-initialised from the base embedder, are added for reference and sink latents. Third, RoPE temporal indices are decoupled from position in the token sequence, allowing references and the lookahead sink to be placed outside the generation window, and attention masks are added for reference and sink padding. All remaining blocks are inherited unchanged.

Chunk geometry. Each chunk covers 77 pixel frames, which the causal VAE encodes to $L = 20$ latents, preceded by $H = 5$ history latents, for a 25-latent generation window. We use $K = 5$ references, a reference gap of $G_{\mathrm{ref}} = 50$ latent positions, an inter-reference spacing of 1, and a lookahead offset of $\Delta_{\mathrm{VL}} = 5$ latent frames. Clean conditioning latents (history, references, and sink) are assigned timestep 0 under the base model's per-frame AdaLN.

Training. We fine-tune from the public Cosmos-Predict2.5-2B checkpoint for 10K iterations at 704×1280 resolution with AdamW ($\mathrm{lr} = 4.8 \times 10^{-5}$, weight decay 0.001, $\beta = (0.9, 0.999)$), 100 warm-up steps, and gradient clipping at 0.1, under the base rectified-flow recipe (logit-normal timestep sampling, shift 5). Training is sharded with FSDP across 8 NVIDIA H100 GPUs at a per-GPU batch size of 1. To avoid over-reliance on any single conditioning signal, we perturb history latents with Gaussian noise of standard deviation 0.1 with probability 0.5, and drop the reference and warped-video pathways independently with probability 0.2 each.

Inference. Chunks are rolled out sequentially with a rectified-flow ODE sampler using 35 steps and classifier-free guidance of 3.0, with only the history latents carried forward; the camera trajectory, retrieved references, and sink are re-derived at every chunk.

A.3 Momentum Implementation Details

WorldUpdate: Multi-Rig DGWM. The rig model is initialised from the released 2B multi-view Cosmos-Predict2.5 checkpoint, with an eighth view-embedding row appended to its seven-slot table and the warped-video latent appended to the input embedder, giving 41 input channels (16 noisy, 1 condition mask, 1 padding mask, 7 view, 16 warped). Every pretrained column keeps its index, so the conversion only adds zero-initialised weights. All eight views of a rig are generated jointly at 192×352 per view with 8 latent frames per view, corresponding to 29 pixel frames per chunk. Generation is conditioned on a single history latent ($h = 1$), with the initial chunk conditioned on one ground-truth frame per view of each rig.

We fine-tune for 10K iterations with AdamW ($\mathrm{lr} = 3 \times 10^{-5}$, weight decay 10^{-3}, $\beta = (0.9, 0.999)$), 100 warm-up steps, and gradient clipping at 0.1, sharded with FSDP. The warped latent is dropped with probability 0.2 per sample, which both teaches the model to generate without a reprojection source and provides the branch without reprojection conditioning for classifier-free guidance. This rate is chosen to approximate how often a rig is the graph source at inference.

WorldInteract: Dynamically Grounded World Model. The target-camera model is fine-tuned from the statically grounded checkpoint described in App. A.2 on the dynamic grounding sessions, retaining that model's chunk geometry, optimiser settings, and conditioning dropout. It generates target-camera video at 704×1280 resolution in chunks of 77 pixel frames ($L = 20$ latents), conditioned on $H = 5$ history latents. At inference, it uses the same 35-step rectified-flow ODE sampler and classifier-free guidance of 3.0 as the statically grounded model.

References are bound to the first frame of the chunk's window. The sink is read $\Delta_{\mathrm{VL}} = 5$ latent frames past the target end, clamped to the length of the available state. During training, a sample's geometric pathway is disabled unless at least 95% of its target frames fall within a window in which the source rig was actively rendering.

Rig connectivity and source selection. Reference rigs actively render within a range of $r = 25\,\text{m}$. This range is recorded with each session and also determines the edges of the rig graph used by WorldUpdate. For target-camera generation, WorldInteract reprojects the nearest in-range rig by default. Rigs outside the active rendering range hold their last frame and are excluded as reprojection sources.

B METRICS

B.1 CAMERA TRAJECTORY ACCURACY

RotErr & TransErr These measure how well the camera motion of the generated video follows the target trajectory, using the protocol of Seo et al. (2026) (App. B.1). We estimate per-frame camera-to-world poses $(\mathbf{R}_t, \mathbf{t}_t)$ of *both* the generated and the ground-truth video with Depth Anything V3 Lin et al. (2025) at a processing resolution of $504\,\text{px}$. Estimating the reference trajectory with the same model, rather than using the simulator's exact poses, reduces the estimator's systematic bias and places both trajectories in the same camera convention. As in Seo et al. (2026), videos are processed in non-overlapping chunks of 32 frames, and each chunk is scored in its own coordinate frame. Within a chunk c with first frame t_0, poses are expressed relative to that anchor,

$$\mathbf{R}_t^{\text{rel}} = \mathbf{R}_{t_0}^\top \mathbf{R}_t, \qquad \mathbf{t}_t^{\text{rel}} = \mathbf{R}_{t_0}^\top \big(\mathbf{t}_t - \mathbf{t}_{t_0}\big), \tag{1}$$

and, since monocular pose estimation is only defined up to scale, each translation trajectory is normalised by its own maximum norm within the chunk, $\bar{\mathbf{t}}_t = \mathbf{t}_t^{\text{rel}} / \max_{t' \in c} \|\mathbf{t}_{t'}^{\text{rel}}\|_2$. With $\mathcal{F}$ the set of all non-anchor frames over all chunks of a clip, the rotation error is the mean geodesic distance on $\mathrm{SO}(3)$ and the translation error the mean Euclidean distance between normalised translations:

$$\text{RotErr} = \frac{1}{|\mathcal{F}|} \sum_{t \in \mathcal{F}} \arccos\left(\frac{\text{tr}\big(\hat{\mathbf{R}}_t^{\text{rel}\,\top} \mathbf{R}_t^{\text{rel}}\big) - 1}{2} \right), \tag{2}$$

$$\text{TransErr} = \frac{1}{|\mathcal{F}|} \sum_{t \in \mathcal{F}} \big\| \hat{\bar{\mathbf{t}}}_t - \bar{\mathbf{t}}_t \big\|_2. \tag{3}$$

The anchor frames, whose error is zero by construction, are excluded. RotErr is reported in radians and TransErr is unitless (it is relative to the chunk's extent); lower is better for both.

B.2 CONSISTENCY METRICS

SCS static & dynamic The Structural Consistency Score Bagchi et al. (2026) measures whether the generated video keeps the *layout* of the scene, independent of appearance, as the mean mask IoU between corresponding regions of the generated and ground-truth videos:

$$\text{SCS} = \frac{1}{NT} \sum_{j=1}^{N} \sum_{t=1}^{T} \frac{\big|\hat{M}^{t,j} \cap M^{t,j}\big|}{\big|\hat{M}^{t,j} \cup M^{t,j}\big|}. \tag{4}$$

Masks are generated automatically with SAM 3.1 Carion et al. (2025) tracking over a fixed vocabulary of semantic categories, and $M^{t,j}$ is the union of all instances of category j tracked in frame t. The static and dynamic categories are:

- **SCS-static**: {*building*, *road*, *sidewalk*, *vegetation*, *sky*}, the persistent scene structure, which the model must preserve and move consistently with the camera;
- **SCS-dynamic**: {*vehicle*, *pedestrian*, *person*, *bicycle*, *animal*}, moving entities, which additionally test whether the model keeps them present and in the right place.

Frames in which a category is empty in *both* videos (IoU $0/0$) are excluded. IoUs are pooled over all valid (clip, category, frame) triples. To bound the cost of SAM 3.1, both videos are temporally subsampled to 2 fps.

B.3 Video Fidelity Metrics

mPSNR & mLPIPS Following Seo et al. (2026), we measure reconstruction fidelity only on the *static* part of the scene. We reuse SAM 3.1 to extract dynamic masks, as above, for the generated and the ground-truth video, giving $\hat{D}_t, D_t \in \{0,1\}^{H\times W}$. A pixel counts as static only if it is static in both videos:

$$S_t = \neg\big(\hat{D}_t \vee D_t\big). \tag{5}$$

mPSNR is the PSNR over all static pixels and colour channels of a clip (peak value 1):

$$\mathrm{mPSNR} = -10\log_{10}\frac{\sum_t\sum_p S_t(p)\,\|\hat{I}_t(p) - I_t(p)\|_2^2}{3\sum_t\sum_p S_t(p)}. \tag{6}$$

LPIPS Zhang et al. (2018) is a distance in feature space and cannot be restricted to an arbitrary pixel set, thus for mLPIPS we fill the dynamic regions of the generated frame with the ground truth, $\tilde{I}_t = S_t \odot \hat{I}_t + (1 - S_t) \odot I_t$, and compute

$$\mathrm{mLPIPS} = \frac{1}{|\mathcal{T}_S|}\sum_{t\in\mathcal{T}_S}\mathrm{LPIPS}_{\mathrm{VGG}}\big(\tilde{I}_t, I_t\big), \qquad \mathcal{T}_S = \{t : \textstyle\sum_p S_t(p) > 0\}. \tag{7}$$

B.4 Measures of Change

Staticness measures *how much of the scene moves* in a video. The exo rig cameras are fixed, so a change in where an object appears in the image can only come from the object moving. We reuse the SAM 3.1 tracks of the dynamic categories. Unlike SCS, masks are compared *across time within one video* rather than between the generated and ground-truth videos. For a clip v, the staticness is the mean IoU between consecutive frames:

$$\mathrm{St}(v) = \frac{1}{|\mathcal{P}_v|}\sum_{(t,j)\in\mathcal{P}_v}\frac{\big|M^{t,j}\cap M^{t+1,j}\big|}{\big|M^{t,j}\cup M^{t+1,j}\big|}, \qquad \mathcal{P}_v = \big\{(t,j) : \big|M^{t,j}\cup M^{t+1,j}\big| > 0\big\}. \tag{8}$$

If no agents are present or moving, $\mathrm{St} = 1$. The score falls as a larger share of the agents' area moves, appears or disappears between frames. As in SCS, both videos are temporally subsampled to 2 fps, so t and $t{+}1$ are 0.5 s apart. Pairs in which a category is empty in both frames are excluded.

We compute St for the generated video ($\widehat{\mathrm{St}}$) and for the ground truth (St), and average each over clips. SAM 3.1 masks jitter slightly from frame to frame, so even a perfectly frozen video scores somewhat below 1; the persistent baseline value is therefore the reference, not 1. A model can match the average staticness while freezing dynamic clips and animating static ones, so we also report the mean per-clip gap

$$\Delta\mathrm{St} = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\big|\widehat{\mathrm{St}}(v) - \mathrm{St}(v)\big|, \tag{9}$$

where $\mathcal{V}$ is the set of clips for which both scores are defined.

C Additional Experiments

C.1 WorldUpdate Consistency

In this experiment we evaluate the multi-view abilities of our WorldUpdate model for generating video across rigs and views. To this end, we take the two closest to the current camera orientations of the neighboring rigs and reproject them to the current view. We evaluate how close the model predictions are to the ground truth of those views at that time. We compare our multi-view WorldUpdate with WorldInteract, which is dynamically grounded but not multi-view in design.

Results are shown in Tab. 5. A baseline that freezes the conditioning frame achieves the best PSNR and LPIPS. This is a property of the setting: the rig cameras are stationary, so most of the image is static in time and visual metrics reward generating no motion. SCS-dyn is affected in the same way. It is an overlap score, and over one evaluation step a moving object still overlaps most of

Table 5: **WorldUpdate consistency across different model choices.** Models are conditioned with ground truth neighbor views reprojections. SCS denotes the Scene Consistency Score, measuring static (SCS-st) and dynamic (SCS-dyn) scene content consistency.

Method	PSNR ↑	LPIPS ↓	SCS-static ↑	SCS-dyn ↑	Stat ↓	Stat Gap ↓
test 1: seen towns						
Persistent Baseline	**33.66**	**0.118**	0.807	**0.208**	0.993	0.565
Dynamically Grounded WM	30.96	0.166	0.800	0.193	0.870	0.480
Multi-Rig DGWM	30.06	0.191	**0.841**	0.124	**0.498**	**0.204**
test 2: unseen town						
Persistent Baseline	**31.88**	**0.110**	**0.801**	**0.228**	0.985	0.593
Dynamically Grounded WM	27.44	0.184	0.723	0.186	0.914	0.519
Multi-Rig DGWM	25.99	0.214	0.730	0.097	**0.427**	**0.234**

itself, so a frozen frame is credited with the overlap the object retains, while a model that moves it is penalised wherever its motion departs from the single recorded future. Both must therefore be read together with the staticness gap. WorldInteract stays close to the frozen baseline in staticness, and inherits its PSNR, LPIPS and SCS-dyn from that. WorldUpdate has the lowest staticness gap in every configuration, moving the scene by roughly the amount the ground truth does, and pays for it on every metric that rewards freezing, while preserving static structure best of all generative models (SCS-st) and exceeding the frozen baseline on it in the seen towns. This experiment thus exposes a trade-off between a preference for static content and dynamic accuracy across models, and a bias of the metrics themselves towards the former.