ARTEMIS: GEOMETRY-GROUNDED MULTI-AGENT DRIVING WORLD MODELS WITH SHARED 3D STATE AND PROGRESSIVE MEMORY UPDATE

Anonymous authors
Paper under double-blind review

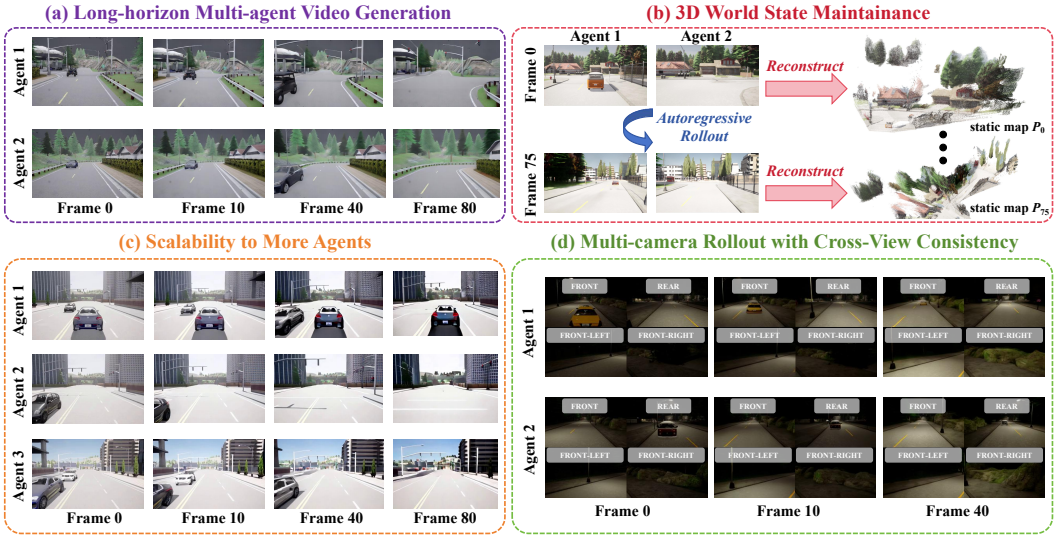


Figure 1: **Artemis: a geometry-grounded multi-agent world model.** (a) Our method can generate long-horizon future videos given initial observations and following actions. (b) Distinct from existing multi-agent methods, we explicitly construct and dynamically update a 3D world state using static point map. (c) With our curated MA-CARLA dataset consisting of various interaction modes, Artemis supports scalable agent numbers beyond two. (d) Furthermore, due to our designed explicit 3D constraints, multi-camera videos can also be created with high cross-view consistency.

ABSTRACT

Recent video world models have witnessed the paradigm shift from single-agent to multi-agent involvements, which can reveal more complicated dynamics and cross-agent interaction in the real world. However, existing approaches commonly adopt implicit inter-agent communications via cross attention, which lack explicit geometry constraints and unified 3D state, thereby leading to poor multi-view consistency and struggling with recovering out-of-sight agents. In addition, most of them assume a static background, failing to represent uncontrolled background dynamics. To address these problems, we propose Artemis: a geometry-grounded multi-agent world model with explicit memory sharing. An explicit 3D world map is reconstructed from multi-agent observations to enforce a unified 3D state across agents, offering high cross-view consistency. Specifically, an action-guided geometric injection module is developed to simultaneously render decomposed foreground-background control maps, which are then injected into a diffusion transformer through a designed GeoAdapter block. Compared to previous methods assuming static-only background, our GeoAdapter can also distinguish uncontrolled non-agent dynamics, which are conditioned on their own multi-frame history positions to provide consistent motion cues. Keyframes selected from progressive video rollouts are used to progressively update the reconstructed 3D world maps. To effectively capture complex dynamic patterns, we curate a novel

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.

dataset sampled from the CARLA simulator called MA-CARLA, with scalable agent numbers, physical plausibility, and abundant interaction modes. Extensive experiments demonstrate the superiority of our proposed method in terms of visual fidelity and cross-view consistency in the generated videos. In addition, our Artemis can support simultaneous multi-modal rollouts with both 2D video and 3D point map maintenance, scale to scenarios beyond two agents, and flexibly switch between single-camera or multi-camera setting.

1 INTRODUCTION

Video world models (Ha & Schmidhuber, 2018; Po et al., 2025; Wu et al., 2025; Liu et al., 2025; Zhao et al., 2025) can predict future environmental dynamics given historical observations and actions from users or agents, which have recently achieved remarkable success in autonomous driving (Wang et al., 2024; Zhao et al., 2025; Liu et al., 2026c;b), embodied AI (Lu et al., 2025; Seo et al., 2023; Huang et al., 2026b), and immersive game engines (Yu et al., 2025; Zhang et al., 2025). However, early methods are mostly constrained by a single-agent perspective, which cannot reflect the intrinsic multi-agent interactions and complex dynamics in the real world.

To effectively extend to multi-agent interactive world models, some recent approaches attempt to achieve inter-agent information exchange in a shared world. As a pioneering work, Solaris (Savva et al., 2026) curates a novel multiplayer dataset in gaming scenes and introduces shared self-attention layers for cross-agent correlation. Subsequent methods further improve it with scalability in agent numbers (Liu et al., 2026a; Wu et al., 2026), accurate agent positional encoding (Sun et al., 2026; Liu et al., 2026a), and world state alignment (Wu et al., 2026; Hu et al., 2026).

However, these multi-agent approaches commonly struggle with three challenges: (1) **Lack of explicit geometric constraint for cross-view consistency**: The physical environment and action-triggered events should evolve consistently and be faithfully reflected from all agents' perspectives. This requires a globally consistent understanding of geometric structures and visual appearances across all agents. MultiWorld (Wu et al., 2026) proposes an action attention and reweighting module to maintain a global state. However, their inter-agent association is still established on implicit cross attention, lacking sufficient geometric cues to ensure multi-view consistency. (2) **Limited memory sharing**: Basically, diverse agents execute their actions and exist in the same shared world, which means that historical observed frames across agents may share high overlap, especially in scenarios when another agent revisits the same location. Therefore, sufficient historical memory sharing is crucial to further strengthen a unified world state and facilitate cross-agent information reuse. However, existing methods (Savva et al., 2026; Wu et al., 2026; Hu et al., 2026) mostly fail to consider effective cross-agent geometry sharing beyond the current timestamp. (3) **Over-simplified scene complexity and dynamics**: Previous methods typically evaluate the multi-agent interaction in the synthetic gameplay scenario where dynamics are quite simplified to reveal the true world evolution, leading to uncertain generalization capabilities in scaling agent complexity, memorizing out-of-sight dynamics, and handling severe occlusions in complex world distributions.

To address these bottlenecks, we propose Artemis, a geometry-grounded multi-agent world model with shared explicit 3D state and memory. Rather than implicit-only agent association, our method constructs a 3D global point cloud-based world map with cross-view consistency to address the first challenge. Explicit geometric constraints are then rendered as control maps and then injected into Diffusion Transformer (DiT) through GeoAdapter, enabling global-aware and view-invariant video generation for each agent. Specifically, we decompose the scene into static background, action-controlled foreground agents, and uncontrolled background dynamics. Static geometry and controlled agents' proxies provide background and foreground RGB and depth control maps, respectively, while uncontrolled non-agent dynamics further require multi-frame historical control maps providing consistent motion tendency for future synthesis.

For the second bottleneck, we further introduce two novel designs to facilitate effective cross-agent memory sharing: (1) *World map sharing.* We progressively select keyframes revealing significantly unobserved areas from new generated videos across all agents to update a global world map. In this case, different agents can explicit share a consistent 3D world state with multi-view matching priors. (2) *Relative pose sharing.* During autoregressive rollout, persistent agent proxies are propagated using cross-agent relative poses and updated from rolling-out videos. Another benefit of

maintaining relative pose is that, when an unrepresented controlled agent is expected to enter the view, a pose-dependent visibility test can initialize its proxy. To address the third challenge, we curate a novel dataset from the CARLA simulator in complex urban scenarios called MA-CARLA, which provides scalable multi-camera sampling, abundant interaction modes, physical plausibility, and diverse environmental dynamics.

Overall, the key contributions of our paper are as follows:

- We propose a novel multi-agent world model called Artemis with strong explicit geometric constraints and cross-agent memory sharing. Compared to prior methods that only leverage implicit feature exchange, our method recovers and dynamically maintains synchronized 3D world states, offering superior cross-view consistency.
- Hybrid geometry-pose memory sharing provides complementary memories for disentangled statics and dynamics. An explicit world map memory is designed by progressively reconstructing static background from new generated videos which are integrated across agents. Additionally, relative-pose visibility checks can support proxy initialization of suddenly emerging vehicles.
- A novel dataset is designed based on the CARLA simulator, which enables physical plausibility, diverse inter-agent interaction modes, multi-camera sampling, and complicated environmental dynamics.
- Extensive experiments demonstrate that our Artemis surpasses prior methods in terms of visual fidelity, cross-view consistency, and uncontrolled background dynamics. Notably, our Artemis can support both single-camera and multi-camera setting, and also simultaneously roll out multi-modal driving scenes with consistent 2D and 3D representations.

2 Related Work

Interactive Video World Models. World models (Ha & Schmidhuber, 2018) can predict future environmental dynamics given historical observations and actions. With rapid advances in recent video diffusion transformers, there is increasing research interest resorting to action-conditioned video generation backbones (Wan et al., 2025) for world modeling. Recently, interactivity has become a crucial property in world models where users can immersively navigate, manipulate, and personalize their created worlds in a multi-round paradigm. As a milestone, Genie (Bruce et al., 2024) compresses the video features into latent space with VAE and then learn actions unsupervisedly, supporting minute-level video generation and interactive exploration. Subsequent world models focus on achieving long-horizon consistency (Po et al., 2025; Huang et al., 2026a; Wang et al., 2026), effective memory design (Wu et al., 2025; Xiao et al., 2025; Hong et al., 2025; Chen et al., 2026a), or real-time rollout (Chen et al., 2026c; Huang et al., 2025; Zhu et al., 2026a). However, most of video world models enforce a single-agent formulation with a first-person or third-person perspective, with limited agent numbers and exploration flexibility.

Multi-Agent Video World Models. To capture inter-instance interactions, MultiVerse (team, 2025) proposes the first multiplayer gaming world model by trivially concatenating each agent's video and action tokens along the channel dimension for denoising together. Solaris (Savva et al., 2026) further designs a shared self-attention and per-player ID embeddings. Nevertheless, its per-slot embeddings break the permutation-symmetry requirement for each agent. Moreover, Solaris suffers from poor generalization to arbitrary agent numbers due to quadratic-complexity full attention. To address these problems, Gamma-World (Liu et al., 2026a) introduces a regular simplex in rotary angle space to achieve equidistant agent embeddings and sparse hub attention to decrease computational costs. MultiWorld (Wu et al., 2026) constructs the testbed on the ItTakesTwo game and develops adaptive weighting in multi-agent action space. To strengthen view consistency, it also introduces VGGT (Wang et al., 2025) as reconstruction priors to represent the 3D-aware global state. MetaWorld (Hu et al., 2026) observes that the collection of multi-view data is prohibitively expensive, thereby decomposing the camera ego-motion and object trajectories for data scaling. MultiGen (Po et al., 2026) breaks down the function in diffusion game engines into memory module, observation module, and dynamic module. Similar to our insight, Prisma-World (Sun et al., 2026) also views scene geometry as significant cues in 3D-aware consistent video generation. However, it only injects the relative camera pose into DiT for geometrical awareness, without explicit 3D maps.



Figure 2: **MA-CARLA dataset.** We visualize various curated samples in our MA-CARLA dataset, and also calculate the distributions of diverse interaction modes and agent numbers.

3 MA-CARLA Dataset

3.1 Data Collection

We construct **MA-CARLA**, a synchronized multi-agent driving dataset rendered in CARLA Dosovitskiy et al. (2017). It contains 11,283 rendered sequences derived from 2,478 unique synchronized trajectories across eight towns and 17 weather presets. The dataset includes 6,020 two-agent sequences with static backgrounds, 3,382 two-agent sequences with dynamic backgrounds containing non-agent actors, and 1,881 three-agent sequences with static backgrounds.

Candidate trajectories are generated using rule-based templates for opposing, following, and overtaking configurations, with route variants that may include turns. After trajectory-level filtering, accepted motions are replayed under different weather conditions while preserving agent motion and synchronization. Each rendered sequence contains 256 synchronized timestamps. Each vehicle is configured with four selected views—front, front-left, front-right, and rear—at 800×600 resolution, together with motion annotations and camera calibration. Across the two-agent sequences, MA-CARLA provides approximately 19.3M available RGB images. Fig. 2 visualizes representative samples and summarizes post-hoc motion groups extracted from recorded trajectories. A sequence may belong to multiple motion groups, yielding 17,790 group assignments in total. Fig. 2 reports their distribution by motion group and agent count. Appendix A.1 provides further details on data collection, grouping criteria, camera configurations, and data splits.

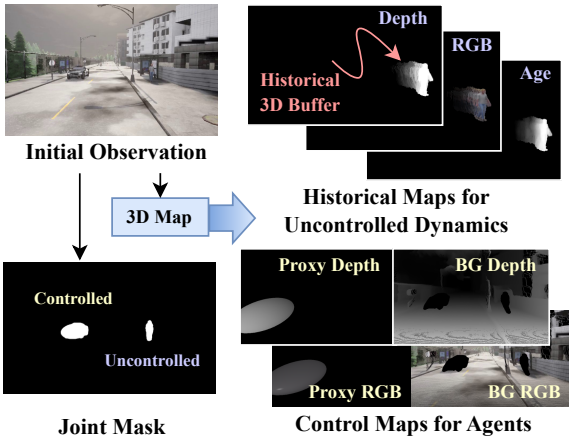


Figure 3: Construction of control maps.

3.2 Data Preprocessing

We preprocess synchronized CARLA videos into geometry-aware conditioning signals. For each agent–camera streaming sequence, LingBot-Map (Chen et al., 2026b) is used to estimates dense depth, intrinsics, and camera poses in a consistent local coordinate system. The same estimates are used for point lifting, proxy fitting, and control rendering to preserve the non-metric scale. Grounded-SAM2 (Ravi et al., 2025; Liu et al., 2024) extracts candidate dynamic masks, associated with either action-controlled agents or uncontrolled actors using CARLA metadata and calibrated projections. We exclude both categories from static-map fusion. Controlled agents are fitted as full-covariance 3D Gaussian proxies, while uncontrolled actors are stored as time-stamped observations rather than assigned future trajectories as in Fig. 3. Artemis then

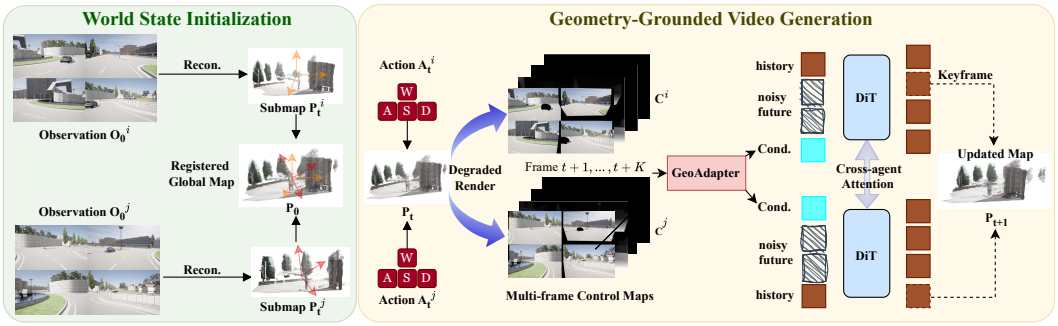


Figure 4: **Overall pipeline of Artemis.** We initialize the shared 3D map by registering individual observations. Given action inputs, control maps disentangling background static, foreground controlled agent proxy, and uncontrolled environmental dynamics are rendered respectively from the reconstructed global map. Then, these controls are injected into DiT, guiding the generator's rollouts. After each inference block, keyframes are selected to update the static world map.

renders a unified 24-channel control tensor consisting of background RGB/depth, proxy RGB/depth, a synthesis mask, uncontrolled-history RGB/depth, and an age mask within a certain historical window.

4 ARTEMIS

4.1 PRELIMINARIES

Problem Formulation. We investigate multi-agent interactive world model in an autoregressive geometry-anchored video generation paradigm which can model joint future probabilistic distributions given sequences of history observations and actions for each agent as:

$$\mathcal{D}_\theta \left(\{\mathbf{F}_t^n\}_{n=1}^N, \mathbf{P}_t | \{\mathbf{O}_{<t}^n, \mathbf{A}_{\leq t}^n\}_{n=1}^N \right), \tag{1}$$

where $\mathbf{O}_t^n$ indicates the history observation and $\mathbf{A}_t^n$ denotes the input action at timestamp t for the agent n. Past observations (memory) together with input actions will jointly formulate the conditional guidance for the generator $\mathcal{D}_\theta$, generating future frames $\mathbf{F}_t^1, ..., \mathbf{F}_t^N$. Here, N is the number of controllable agents, and other moving scene objects are treated as uncontrolled environmental dynamics. Previous multi-agent methods (Sun et al., 2026; Savva et al., 2026; Wu et al., 2026; Hu et al., 2026) use primarily 2D-based world representations, suffering from degraded cross-camera and cross-agent consistency and lacking 3D geometric guidance. To tackle these problems, we reconstruct and progressively update a sparse point cloud $\mathbf{P}_t$ as a shared 3D state for all agents. Furthermore, our Artemis also supports multi-camera driving data through camera tiling (Zhu et al., 2026b), which enlarges sensor coverage for real-world applications and mitigates the potential occlusions in single-camera-only approaches.

Overall Pipeline. As in Fig. 4, we first initialize a shared static world map from the agents' initial observations with LingBot-Map (Chen et al., 2026b), a feed-forward foundation model for streaming 3D reconstruction, where submaps from all agents are registered in a common coordinate system. Given the input actions, we derive agent trajectories and render foreground-background disentangled control maps from the reconstructed world (Yang et al., 2026). Specifically, controlled-agent proxies fitted from the first frames are propagated along these trajectories, with visibility-triggered agent insertion when needed. We render the static map, controlled proxies, and available uncontrolled dynamics observation history into a unified set of controls for the planned camera views, encoding both geometry and historical dynamics. Afterwards, a GeoAdapter module injects these controls into a pre-trained video diffusion transformer, Wan2.2 5B (Wan et al., 2025), which jointly generates synchronized future frames for all agents. Another cross attention layer (Savva et al., 2026; Wu et al., 2026) is also designed to communicate inter-agent information in latent space. After generation, reconstructed static content updates the shared map, and accepted uncontrolled dynamics update their 3D history. Clean generated latents also update the implicit per-agent historical memories in KV-Cache for autoregressive online rollout.

4.2 SHARED 3D WORLD STATE CONSTRUCTION

In real world, various agents perceive and interact with the same evolving world, and the observations across agents are just different projections of the shared 3D geometry, each captured from a specific viewpoint. Motivated by this insight, we explicitly reconstruct a unified 3D world state from agents' observations and actions, ensuring that the generated videos can maintain high cross-view consistency and spatial awareness.

4.2.1 WORLD INITIALIZATION

Given the first-frame single- or multi-camera observations from all agents $\{\mathbf{O}_0^n\}_{n=1}^N$, we initialize the world map using a streaming feed-forward reconstruction model (Chen et al., 2026b). Specifically, for each agent n, initial submaps $\mathbf{P}_0^n$ represented by static point clouds are reconstructed from their individually observed first frames and merged together to construct the shared static global map $\mathbf{P}_0 = \cup_{n=1}^N(\mathbf{P}_0^n)$ using the cross-agent relative pose. We exclude controlled agents, uncontrolled environmental dynamics, together with sky regions and low-confidence geometry, from static-map fusion. Furthermore, we also maintain an individual history 3D buffer $\mathcal{H}^n$ for each segmented uncontrolled dynamics to store their motion priors.

4.2.2 ACTION-GUIDED GEOMETRIC INJECTION

Future videos are jointly conditioned by historical observations and the input actions of each agent. To effectively control agent trajectories, we propose an action-guided geometric injection module consisting of three components: Action-conditioned Trajectory Sampling, Foreground-background Disentangled Control Signal Rendering, and GeoAdapter for latent-space geometry injection.

Action-conditioned Trajectory Sampling. During online inference, discrete action inputs can determine the future motion of the controllable agents within each rollout block. For each controllable agent n, an action $\mathbf{a}^n$ is represented as a local transformation $\Delta\mathbf{T}(\mathbf{a}^n) \in \mathrm{SE}(3)$ relative to its current camera-to-world pose $\mathbf{T}_t^n$. To smoothly recover motions at any intermediate timestamp, we interpolate a sequence of intermediate trajectories from the current to the target pose $\mathbf{T}_t^n\Delta\mathbf{T}(\mathbf{a}^n)$. over K timestamps, obtaining $\mathbf{T}_t^n, \mathbf{T}_{t+1}^n, \ldots, \mathbf{T}_{t+K}^n$. Translation is linearly interpolated, while rotation is interpolated on $\mathrm{SO}(3)$.

Foreground-background Disentangled Rendering. We construct the per-block control maps by independently rendering static background geometry, action-controlled foreground proxies, and history-conditioned control maps for uncontrolled dynamics. For each future pose $\mathbf{T}_{t+k}^n$ of the observer agent n, the explicit world state $\mathbf{P}_t$ is rendered to obtain a degraded background RGB image, a background depth map, and a visibility mask, while controllable foreground proxies are rendered to obtain foreground degraded RGB, depth, and visibility mask as:

$$\mathbf{B}_{\mathrm{rgb}}^n, \mathbf{B}_{\mathrm{dep}}^n, \mathbf{M}_{\mathrm{hole},k}^n = \mathcal{R}_{\mathrm{bg}}\left(\mathbf{P}_t, \mathbf{T}_{t+k}^n, \mathbf{I}^n\right), \quad \mathbf{F}_{\mathrm{rgb}}^n, \mathbf{F}_{\mathrm{dep}}^n, \mathbf{M}_{\mathrm{ob},k}^n = \mathcal{R}_{\mathrm{ob}}\left(\mathcal{Q}_{t+k}^{\neg n}, \mathbf{T}_{t+k}^n, \mathbf{I}^n\right), \tag{2}$$

where $\mathbf{I}^n$ indicates the camera intrinsics. The background-hole mask $\mathbf{M}_{\mathrm{hole},k}^n$ and foreground mask $\mathbf{M}_{\mathrm{ob},k}^n$ identifies regions that cannot be rendered by the current 3D map due to missing geometry, occlusion, or sparsity. $\mathcal{Q}_{t+k}^{\neg n}$ denotes the proxy set.

For uncontrolled dynamics, let $\mathcal{H}^n$ denote the historical 3D buffer available to observer n. At target frame $t+k$, we select observations from the buffer within a temporal window: $\mathcal{S}(t+k;a) = \{s : \max(0, t+k-L+1) \le s \le \min(t+k, a)\}$, where $a = t+k$ for offline preprocessing and a is the next rollout block boundary during online inference. By reprojecting these observations, we obtain:

$$\mathbf{H}_{\mathrm{rgb}}^n, \mathbf{H}_{\mathrm{dep}}^n, \mathbf{H}_{\mathrm{age}}^n = \mathcal{R}_{\mathrm{hist}}\left(\mathcal{H}_{\mathcal{S}(t+k;a)}^n, \mathbf{T}_{t+k}^n, \mathbf{I}^n\right), \tag{3}$$

where $\mathbf{H}_{\mathrm{age}}^n$ is an age-weighted soft coverage map. Finally, we combine the background-hole and foreground masks together: $\mathbf{M}^n = \mathbf{M}_{\mathrm{hole},k}^n \vee \mathbf{M}_{\mathrm{obj},k}^n$. The integrated mask highlights both newly exposed regions and dynamic foreground regions that still require synthesis. Overall, the complete control maps are: $\mathbf{C}^n = \mathrm{Concat}[\mathbf{B}_{\mathrm{rgb}}^n, \mathbf{B}_{\mathrm{dep}}^n, \mathbf{F}_{\mathrm{rgb}}^n, \mathbf{F}_{\mathrm{dep}}^n, \mathbf{M}^n, \mathbf{H}_{\mathrm{rgb}}^n, \mathbf{H}_{\mathrm{dep}}^n, \mathbf{H}_{\mathrm{age}}^n]$, including static background RGB/depth, controllable-agent RGB/depth, the synthesis mask, and uncontrolled-history RGB/depth/age values. As for the multi-camera setting, in addition to RGB/depth controls, we also provide the DiT through token-aligned camera geometry with ray-map.

GeoAdapter. The rendered unified control maps are then injected into latent spaces of DiT through our GeoAdapter branch. Specifically, the RGB, depth and age controls are encoded via VAE, while the visibility mask is directly downsampled using nearest neighbor interpolation. These compressed latent controls are then embedded and injected into selected DiT layers through residual feature projections. Overall, the explicit geometry control maps can determine where proxies should appear, and the pretrained DiT correspondingly imagines the contents' appearance.

4.3 Memory-Enhanced Streaming World Update

To enhance cross-agent consistency and achieve progressive state update, we design three world update mechanisms for static background, uncontrolled dynamics, and agent pose, respectively.

Static Map Update. After each rollout, we will select a few keyframes from newly generated frames in the block to update the unified 3D world map. The chosen keyframes are the frames introducing unseen scenarios or sharing low overlap with historical observations. The final world map is updated by merging the original static geometry map and the newly recovered submap reconstructed from keyframes.

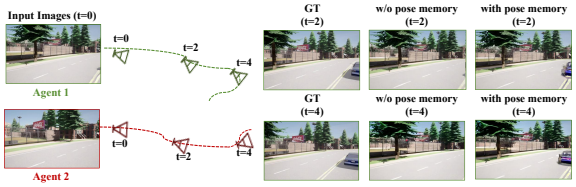


Figure 5: Illustration of our designed pose memory.

Uncontrolled Dynamic Buffer Maintenance. For each uncontrolled segmented dynamics, we update the history 3D buffer using a first-in-first-out principle, discarding the farthest frames while dynamically retaining the latest frame. In this case, the future positions can be extrapolated from the progressively updated historical motion priors.

Structured Pose Memory Maintenance. For each controllable agent, we maintain their historical poses across rollout blocks, so that an agent can remain geometrically consistent even when it is temporarily occluded or outside the observer's sight. If one agent is not visible in the current frame, the generation model cannot generate it in the following block as original control maps fail to aware its existence through current frame segmentation. To address this issue, we design a pose=guided verification module based on maintained pose memory as in Fig. 5. If no proxy is available is the current frame but the target's planned poses indicate potential entry into the observer's field of view, we initialize a conservative vehicle-shaped Gaussian and integrate it with the target position. This procedure allows the model to possess a global awareness of each agent's location and feasibly supports the generation of newly visible controlled agents at any timestamp.

5 Experiment

5.1 Implementation Details

Artemis is built on the Wan2.2-TI2V-5B DiT (Wan et al., 2025) backbone. We use a three-stage training recipe with Diffusion-Forcing (Chen et al., 2024) and Self-Forcing (Huang et al., 2025) tailored for multi-agent online rollout. In Stage 1, we train the bidirectional model on paired CARLA videos with 15-channel geometric control maps using AdamW with an initial learning rate of 4×10^{-5}, 200 warmup steps, and cosine decay on 32 H20 GPUs. In Stage 2, we initialize from the Stage-1 checkpoint and finetune it using AdamW with a learning rate of 2×10^{-5} and 400 warmup steps. Both stages are trained on 81-frame clips at 256×416 resolution with flow matching, 1000 training timesteps, and flow shift $\gamma = 5.0$. During online inference, we generate 81-frame videos block by block with 5-frame denoising blocks and a 33-frame history window, using 50-step Euler sampling, guidance scale $\eta = 1.0$. For stage-3 distillation, a causal generator is optimized with DMD-style distillation with 5-frame blocks and a 2-step denoising schedule.

5.2 Quantitative Results

Comparison Baselines. Since recent multi-agent world models are mostly still close-sourced, we choose a representative method Solaris (Savva et al., 2026), using implicit shared self-attention for cross-agent association. We also migrate a similar strategy as in OASIS (Decart et al., 2024) by direct concatenating multi-agent tokens along feature channels.

Table 1: Quantitative comparisons on MA-CARLA Dataset.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	Subject Consistency ↑	Background Consistency ↑	Aesthetic Quality ↑	Imaging Quality ↑
OASIS	13.459	0.500	0.530	0.828	0.921	0.440	0.434
Solaris	10.900	0.378	0.660	0.759	0.916	0.392	0.525
Solaris (finetuned)	16.015	0.522	0.414	0.896	**0.953**	0.502	**0.581**
Ours	**17.743**	**0.566**	**0.309**	**0.902**	0.949	**0.536**	0.568

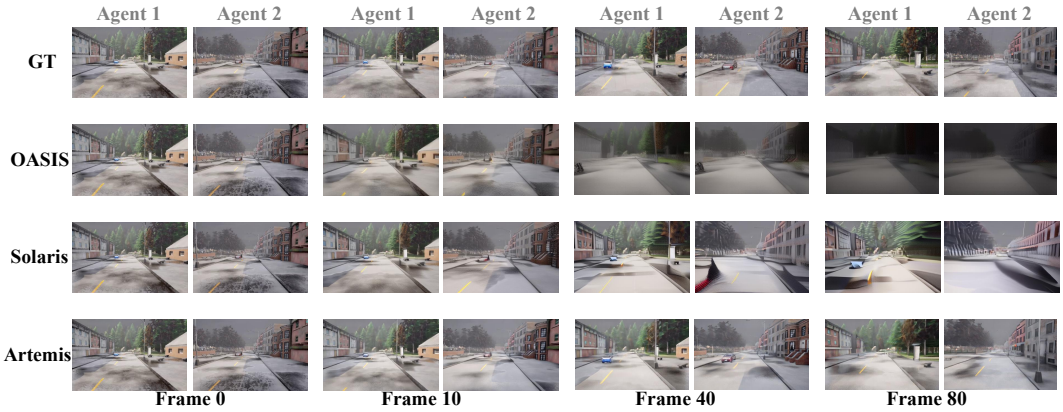


Figure 6: Visualization comparisons with previous multi-agent methods.

Evaluation Metrics. We report eight metrics covering both appearance fidelity and multi-agent consistency. For appearance metrics, we compute PSNR, SSIM, and LPIPS between each generated video and its ground-truth video. For multi-agent consistency, we use DA3 (Lin et al., 2025) to process each generated video and estimate depth, camera intrinsics, and camera poses. Subject Consistency and Background Consistency evaluate the temporal stability of subject appearance and background structure, respectively. We also report the Aesthetic Quality and Imaging Quality, reflecting frame-level aesthetic appeal and overall perceptual quality, respectively.

Comparison Results. As in Table 1, our Artemis outperforms the baseline method Solaris and OASIS by a large margin in both appearance-related metrics and also consistency. Compared to Solaris, our Artemis achieves almost double SSIM, while reducing LPIPS to almost half. This proves that our rendered videos possess higher generation quality with an explicit geometry-guided paradigm and powerful generation priors from DiT. As for multi-agent consistency, our method also significantly surpasses Solaris in both cross-view consistency and trajectory consistency. This is mainly because implicit-only methods cannot recover underlying 3D world states, suffering from severe camera pose changes, occlusions, or few-shot views.

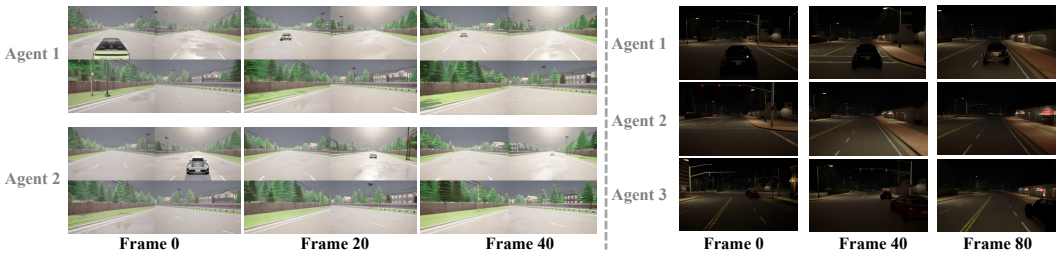
5.3 Qualitative Results



Figure 7: Visualization of uncontrolled dynamics.

Comparison with Implicit Multi-agent World Models. Previous multi-agent world models commonly adopt an implicit agent association manner, where cross-agent attention is introduced to exchange perspectives across agents. However, we observe that this implicit modeling can induce inconsistency geometry, especially for longer-sequence predictions. As in Fig. 6, OASIS cannot handle the driving scenarios, with almost all blurry future imaginations. Although Solaris can generate basically reasonable geometry within the next 10 future frames, it suffers from artifact during long-horizon rollout. This illustrates that Solaris cannot learn genuine geometric consistency over time. In contrast, Artemis still forecasts accurate future videos over long sequences. We also visualize the uncontrolled dynamics in Fig. 7, where two non-rigid bicyclists are predicted accurately.

Multi-camera Rollout. Unlike previous multi-agent world models, our collected data and training pipeline can support multi-camera settings, where more spatial ranges could be considered for better geometric reconstruction. In Fig. 8 (a), we also provide a multi-camera sample where videos



(a) Multi-camera Video Rollout **(b) Multi-agent Video Rollout**

Figure 8: Visualization samples in the multi-camera and multi-agent settings.

Table 2: Ablation studies of our designed modules.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	Subject Consistency ↑	Background Consistency ↑	Aesthetic Quality ↑	Imaging Quality ↑
w/o proxy depth	17.5823	0.5613	0.3161	0.8982	0.9475	0.5352	**0.5707**
w/o proxy RGB	17.3938	0.5555	0.3287	0.8966	0.9471	0.5350	0.5619
w/o ray-map	17.5284	0.5601	0.3263	0.8902	0.9456	0.5327	0.5635
w/o static map update	**17.7429**	**0.5666**	0.3092	0.8998	0.9486	0.5356	0.5671
with separate static map	17.7236	0.5657	0.3093	0.9019	0.9494	0.5358	0.5675
with truncated KV-Cache	17.6488	0.5628	0.3167	0.8926	0.9465	0.5297	0.5562
Ours	**17.7429**	0.5662	**0.3085**	**0.9021**	**0.9495**	**0.5360**	0.5676

samples from the directions `FRONT`, `FRONT_LEFT`, `FRONT_RIGHT`, and `REAR` are delivered as the conditional input. Due to the explicit reconstructed 3D state, our Artemis can still forecast cross-view consistent future videos.

5.4 Scalability to More Agents

For scalability beyond two agents, we extend the pairwise pipeline by treating each vehicle as an observer and all other vehicles as controlled targets. Unlike the two-agent case, where the visible foreground object can usually be assigned to the other agent directly, the multi-agent setting requires identity-aware target matching because each observer may see multiple vehicles. For each observer, detected vehicle masks are matched to the other agents using projected trajectories, proxy-track predictions, depth consistency, and image-space distance. The matched targets are then fitted as proxy Gaussians, rolled forward with their own trajectories, and merged into one dynamic control map for that observer. Online inference results of three agents are shown in Fig. 8 (b).

5.5 Ablation Studies

We conduct ablation studies to analyze the contribution of the geometric conditioning, online world update, and temporal memory mechanisms. The results are summarized in Table 2. First, removing depth-related conditions consistently degrades reconstruction quality. This confirms that the dynamic proxy depth provide complementary geometric guidance. Second, disabling online world map update also hurts performance, reducing background and subject consistencies. This indicates that maintaining and updating the world representation during rollout is important for long-horizon consistency. Without map update, accumulated generation errors cannot be corrected by newly observed content, leading to weaker view reconstruction quality. Reducing the KV-cache with fewer historical frames also leads to worse perceptual quality and consistency. The result suggests that the causal KV-cache is important for preserving coherent autoregressive history.

6 Conclusion

In this paper, we propose a novel multi-agent world model which is anchored by reconstructed 3D world state. To incorporate large-scale dynamics, we curate a new driving dataset based on the CARLA simulator. Unlike previous methods mostly leveraging implicit cross-agent correlation, our Artemis can explicitly reconstruct and dynamically update a synchronized 3D world map. We also design a history 3D buffer for uncontrolled environmental dynamics, leveraging historical observations to predict their future motion. Extensive experiments demonstrate the effectiveness of our proposed novelty, spatio-temporal consistency across frames, and scalability to more agent numbers.

AI USE STATEMENT

LLMs were used to assist with writing and presentation, including language polishing, reorganizing existing manuscript content, and checking formatting and textual consistency. It was not used to run experiments, generate experimental measurements, or determine the reported results. The authors are responsible for the final text, claims, figures, and results.

ETHICS STATEMENT

Artemis is developed for research on multi-agent world modeling and autonomous driving simulation. The proposed method generates future driving videos and maintains a shared 3D world state, but its performance in simulated environments does not establish safety for real-world autonomous-driving deployment. Safety-critical use would require additional system-level validation, robustness evaluation under distribution shifts, and appropriate human oversight. The MA-CARLA dataset used in this work is generated in the CARLA simulator and does not contain personally identifiable information from real-world participants.

REPRODUCIBILITY STATEMENT

Section 3 describes the construction and preprocessing of the MA-CARLA dataset. Section 4 presents the shared 3D world-state construction, action-guided geometric injection, GeoAdapter, and memory-enhanced streaming world update. Section 5.1 reports the backbone, training configuration, inference settings, and three-stage training recipe, while Table 2 provides ablation studies of the main design components. Appendix A.1 further details dataset collection, interaction templates, camera settings, data splits, and scale alignment, and Appendix A.2 describes the three-stage training procedure.

REFERENCES

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 *International Conference on Machine Learning*, 2024.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

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.

Lin-Zhuo Chen, Jian Gao, Yihang Chen, Ka Leong Cheng, Yipengjing Sun, Liangxiao Hu, Nan Xue, Xing Zhu, Yujun Shen, Yao Yao, et al. Geometric context transformer for streaming 3d reconstruction. *arXiv preprint arXiv:2604.14141*, 2026b.

Shuo Chen, Cong Wei, Sun Sun, Ping Nie, Kai Zhou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context forcing: Consistent autoregressive video generation with long context. *arXiv preprint arXiv:2602.06028*, 2026c.

Etched Decart, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer. *URL: https://oasis-model. github. io*, 2(3):6, 2024.

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.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, Sai Bi, Yannick Hold-Geoffroy, Mike Roberts, Matthew Fisher, Eli Shechtman, et al. Relic: Interactive video world model with long-horizon memory. *arXiv preprint arXiv:2512.04040*, 2025.

Teng Hu, Mingchun Lu, Yating Wang, Jiangning Zhang, Jinkun Hao, Ye Pan, Ran Yi, Lizhuang Ma, and Dacheng Tao. Metaworld: Scaling multi-agent video world model from single-view video data. *arXiv preprint arXiv:2606.02753*, 2026.

Junchao Huang, Ziyang Ye, Xinting Hu, Tianyu He, Guiyu Zhang, Shaoshuai Shi, Jiang Bian, and Li Jiang. Live: Long-horizon interactive video world modeling. *arXiv preprint arXiv:2602.03747*, 2026a.

Wenlong Huang, Yu-Wei Chao, Arsalan Mousavian, Ming-Yu Liu, Dieter Fox, Kaichun Mo, and Li Fei-Fei. Pointworld: Scaling 3d world models for in-the-wild robotic manipulation. *arXiv preprint arXiv:2601.03782*, 2026b.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, 2025.

Haotong Lin, Sili Chen, Junhao 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.

Fangfu Liu, Kai He, Tianchang Shen, Tianshi Cao, Sanja Fidler, Yueqi Duan, Jun Gao, Igor Gilitschenski, Zian Wang, and Xuanchi Ren. Gamma-world: Generative multi-agent world modeling beyond two players. *arXiv preprint arXiv:2605.28816*, 2026a.

Mengmeng Liu, Diankun Zhang, Jiuming Liu, Jianfeng Cui, Hongwei Xie, Guang Chen, Hangjun Ye, Francesco Nex, Hao Cheng, and Michael Ying Yang. Universe: Unified video action models for autonomous driving with flexible mask-modulated modality generation. *arXiv preprint arXiv:2607.05133*, 2026b.

Mengmeng Liu, Diankun Zhang, Jiuming Liu, Jianfeng Cui, Hongwei Xie, Guang Chen, Hangjun Ye, Michael Ying Yang, Francesco Nex, and Hao Cheng. Driveva: Video action models are zero-shot drivers. In *European Conference on Computer Vision*, pp. 315–335, 2026c.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55, 2024.

Zhiheng Liu, Xueqing Deng, Shoufa Chen, Angtian Wang, Qiushan Guo, Mingfei Han, Zeyue Xue, Mengzhao Chen, Ping Luo, and Linjie Yang. Worldweaver: Generating long-horizon video worlds via rich perception. In *Advances in Neural Information Processing Systems*, 2025.

Guanxing Lu, Baoxiong Jia, Puhao Li, Yixin Chen, Ziwei Wang, Yansong Tang, and Siyuan Huang. Gwm: Towards scalable gaussian world models for robotic manipulation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9263–9274, 2025.

Ryan Po, Yotam Nitzan, Richard Zhang, Berlin Chen, Tri Dao, Eli Shechtman, Gordon Wetzstein, and Xun Huang. Long-context state-space video world models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8733–8744, 2025.

Ryan Po, David Junhao Zhang, Amir Hertz, Gordon Wetzstein, Neal Wadhwa, and Nataniel Ruiz. Multigen: Level-design for editable multiplayer worlds in diffusion game engines. *arXiv preprint arXiv:2603.06679*, 2026.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Georgy Savva, Oscar Michel, Daohan Lu, Suppakit Waiwitlikhit, Timothy Meehan, Dhairya Mishra, Srivats Poddar, Jack Lu, and Saining Xie. Solaris: Building a multiplayer video world model in minecraft. *arXiv preprint arXiv:2602.22208*, 2026.

Younggyo Seo, Junsu Kim, Stephen James, Kimin Lee, Jinwoo Shin, and Pieter Abbeel. Multi-view masked world models for visual robotic manipulation. In *International Conference on Machine Learning*, pp. 30613–30632, 2023.

Huiqiang Sun, Zhan Peng, Size Wu, Kun Wang, Kang Liao, Dianyi Wang, Xingyu Zeng, Sheng Jin, Yangguang Li, Zhiguo Cao, et al. Prisma-world: Camera-controllable multi-agent video world model. *arXiv preprint arXiv:2606.09507*, 2026.

Enigma team. Introducing multiverse: The first ai multiplayer world model, 2025. URL `https://enigma.inc/blog`.

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

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European conference on computer vision*, pp. 55–72, 2024.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026.

Haoyu Wu, Jiwen Yu, Yingtian Zou, and Xihui Liu. Multiworld: Scalable multi-agent multi-view video world models. *arXiv preprint arXiv:2604.18564*, 2026.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. In *Advances in Neural Information Processing Systems*, 2025.

Zeqi Xiao, Yushi LAN, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. In *Neural Information Processing Systems*, 2025.

Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. Neoverse: Enhancing 4d world model with in-the-wild monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40340–40351, 2026.

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

Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Fei Kang, Biao Jiang, Zedong Gao, Eric Li, Yang Liu, et al. Matrix-game: Interactive world foundation model. *arXiv preprint arXiv:2506.18701*, 2025.

Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In *AAAI Conference on Artificial Intelligence*, volume 39, pp. 10412–10420, 2025.

Haoyi Zhu, Haozhe Liu, Yuyang Zhao, Tian Ye, Junsong Chen, Jincheng Yu, Tong He, Song Han, and Enze Xie. Sana-wm: Efficient minute-scale world modeling with hybrid linear diffusion transformer. *arXiv preprint arXiv:2605.15178*, 2026a.

Jiayi Zhu, Jianing Zhang, Yiying Yang, Wei Cheng, and Xiaoyun Yuan. Shareverse: Multi-agent consistent video generation for shared world modeling. *arXiv preprint arXiv:2603.02697*, 2026b.

A APPENDIX

In the appendix, we supplement the additional implementation settings, data collection and processing procedures, and more experimental results. Overall, the structure is as follows:

- More details about the dataset curation, interaction mode design, data splits, and preprocessing are provided in Section A.1.
- Three-stage training procedures are introduced in Section A.2.
- More visualization results are further provided in Section A.3.

A.1 MA-CARLA DATASET DETAILS

A.1.1 DATA COLLECTION

Each weather rendering of a trajectory counts as a scene, irrespective of camera count. Dynamic backgrounds add moving pedestrians or cyclists, which are not agents. Instrumented vehicles may move under either background setting. Before rendering, we check collisions and spacing, then assess visibility, projected size, and motion smoothness. Accepted trajectories are replayed with synchronized agents and cameras.

A.1.2 INTERACTION TEMPLATE DESIGN

Templates arrange vehicles to approach from opposite directions, follow one another, or overtake in adjacent lanes. Route variants assign turns. The dynamic two-agent pool also includes braking, stop-and-go motion, and lane changes. Initial spacing and speeds vary across routes. Three-agent templates arrange opposing encounters, chases involving a parallel pair, or adjacent-lane overtakes. Each episode can contain several pairwise relationships.

A.1.3 FIGURE STATISTICS

Figure 2 summarizes recorded motion rather than generator labels. Opposing encounters are identified from relative heading and proximity. The co-directional group includes following and parallel travel, with overtaking identified by a reversal of vehicle order. The turning and crossing group covers heading changes, including road bends, or intersections of recorded paths. Spatial crossing does not require simultaneous arrival. We evaluate all vehicle pairs and count each scene once per satisfied group. Groups may overlap. Both rings combine background conditions and use 17,790 group assignments as the denominator for angles and percentages. Outer sectors divide each group by agent count. The center reports the 11,283 rendered scenes.

A.1.4 CAMERAS AND DATA SPLITS

The selected views face front, front-left, front-right, and rear, with fields of view in Table 3. Intrinsics and poses are stored per scene. Recorded timestamps determine temporal spacing, which varies across collection configurations.

Table 3: Horizontal fields of view for the selected cameras.

Configuration	Front	Oblique (each)	Rear
Two-agent static, narrow	70°	70°	90°
Two-agent static, wide	80°	85°	95°
Two-agent dynamic	75°	75°	90°
Three-agent pool	70°	70°	90°

Table 4 reports the experimental splits. Multicamera experiments use the static scenes with complete camera streams. Frozen trajectory groups keep weather re-renders and all observations together within each pool. The three-agent pool is separate from the reported experiments.

Table 4: Scene counts in the two-agent experimental splits.

Setting	Train	Validation	Test
Static, single camera	4,816	602	602
Static, four-view tile	4,813	600	601
Dynamic background	2,706	338	338

A.1.5 SCALE ALIGNMENT

In preprocessing, each agent's observation stream is reconstructed with LingBot-Map, which estimates dense depth, camera intrinsics, and camera poses in a consistent local coordinate gauge. We use these estimates jointly for background point lifting, controllable-proxy fitting, ambient-history lifting, and control-map rendering. CARLA annotations are used for identity association and motion supervision, but any metric quantities used by the renderer are converted into the LingBot-Map gauge before being combined with reconstructed geometry.

During online rollout, the initial observations define the reference scale and camera rig geometry. LingBot-Map reconstructs the initial local scene and provides the intrinsics and camera poses used for subsequent rendering and ray-map construction. User actions, inter-agent offsets, proxy extents, and generated-frame reconstructions are expressed in this same reference gauge before map fusion or dynamic-history insertion. Thus, Artemis maintains a single consistent geometric scale throughout autoregressive rollout, without requiring per-block scale re-estimation.

A.2 THREE-STAGE TRAINING PROCEDURE

We design a three-stage training recipe using Diffusion-Forcing (Chen et al., 2024) and Self-Forcing (Huang et al., 2025) and tailor them for multi-agent online rollout:

Stage 1: Bidirectional Teacher for Geometry Grounding. We first train a bidirectional teacher network to guarantee high-quality geometry reconstruction and effective control map injection. This stage jointly trains the cross-agent attention layer, GeoAdapter, and DiT on full sequences for all agents, using pre-processed control maps. Stage 1 can exploit larger temporal receptive field, cross-agent interactions, and spatio-temporal consistency, which serves as a real teacher for stage 3 distillation.

Stage 2: Causal Student with Diffusion-Forcing. Stage 2 training is initialized based on the pretrained checkpoint form stage 1. However, we switch the temporally non-causal training in stage 1 to autoregressive training with causal masks using Diffusion-Forcing (Chen et al., 2024). Variant noise scales are implemented on different video frames in DiT, while the noise scales for various agents are also independently sampled. Notably, the generation process is still established on full frame sequences in a single forward pass.

Stage 3: Self-Forcing-based Distillation. For stage 3, we leverage online streaming rollout, autoregressively generating videos based on history self-rollouts in a chunk-wise paradigm (Huang et al., 2025), as in Sec. 4.1. This step reduces the diffusion denoising process into fewer steps which improves online inference efficiency. Specifically, control maps are first online rendered from the initialized 3D reconstructed world and intermediate actions. Then, latent frames are generated causally, where each agent will maintain an independent KV Cache to store its own historical latent frames as memory. Finally, the generated chunk is decoded into current videos, which are then used to reconstruct the updated 3D world map. We instantiate three models with identical architectures but distinct roles: a causal generator initialized from stage 2, a frozen bidirectional real teacher initialized from stage 1, and a trainable bidirectional fake critic, also initialized from stage 1. The real teacher represents the data distribution learned from the real videos, whereas the fake critic continuously tracks the evolving distribution induced by the causal generator.

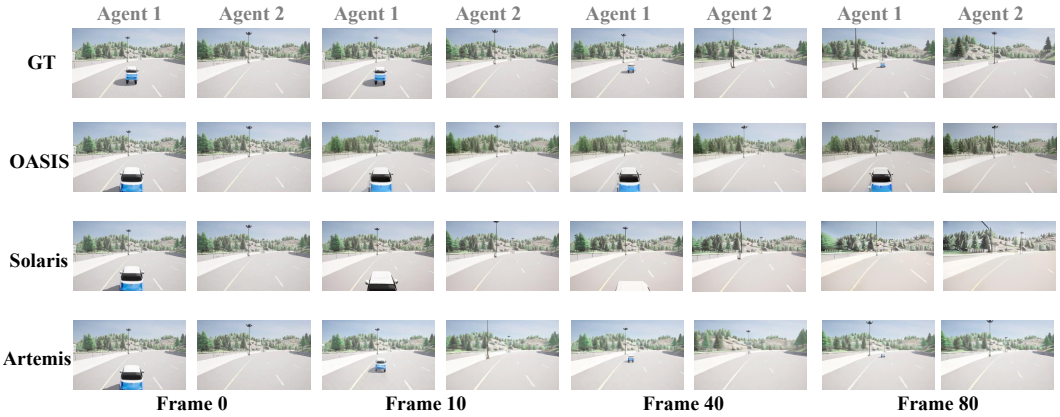


Figure 9: Visualization comparisons with previous multi-agent methods.

A.3 Additional Visualizations

We further visualize another sample in our generated videos in Fig. 9. Previous implicit methods estimate wrong agent motions, where OASIS predicts a static agent while Solaris estimates a reversed direction. Compared to them, our method can estimate accurate agent motions.