WAM-SIM: WORLD ANCHORED MEMORY MEETS ONLINE RL FOR CONSISTENT AND REALISTIC DRIVING SIMULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive driving world models systematically suffer from structural drift, identity drift, and quality decay over long rollout horizons due to transient conditioning signals that lack persistent state. We present WAM-Sim, a driving world model anchored by a persistent hierarchical memory and optimized via online flow-matching reinforcement learning. WAM-Sim structures persistent context across three temporal and semantic scales: static structural tokens encoding road topology, read-only scene tokens supplying geolocated background appearance, and write-once instance tokens preserving individual traffic participant identities across occlusions and chunk boundaries. To restore visual sharpness and physical plausibility degraded by multi-scale conditioning, we introduce an online post-training framework that optimizes self-generated rollouts via flow-matching RL guided by hierarchy-aligned visual rewards and a vision-language model physical judge. The invariant memory manifold bounds policy optimization to suppress semantic collapse and reward hacking, while online RL systematically enforces physical commonsense and fine-grained visual fidelity. On the nuScenes benchmark, WAM-Sim achieves a superior generation fidelity (FID=7.07, FVD=42.01) and strong downstream perception performance (19.67 mAP, 68.08 road mIoU). On minute-scale rollouts of self-collected data, WAM-Sim substantially mitigates structural and identity drift, preserving high scene similarity and subject identity coherence across extended horizons.

1 INTRODUCTION

Video world models offer a scalable paradigm for controllable closed-loop simulation in autonomous driving, synthesizing driving observations conditioned on ego trajectories and agent controls (Hu et al., 2023; Gao et al., 2024c; Wen et al., 2024; Russell et al., 2025). Despite recent progress, standard training and evaluation regimes remain predominantly constrained to short temporal clips and instantaneous point-estimate metrics. As rollouts extend toward minute-scale horizons, generation systematically degrades through three recurring failure modes: 1) structural drift, where road geometry and background appearance detach from geographic reality; 2) identity drift, where traffic participants undergo unconstrained appearance, category, or tracking mutations; and 3) quality decay, where accumulated autoregressive errors recursively compound over time. Although recent architectures incorporate extended context windows, temporal filtering, or instance masking (Guo et al., 2024; Huang et al., 2025b; Liu et al., 2026; Cheng et al., 2026; Basant et al., 2026), these capabilities operate in isolation without a persistent and addressable state. We argue that long-horizon generation failure is not merely a function of limited context window length, but fundamentally demands persistent, multi-scale memory semantics to anchor identity, appearance and geometry.

In conventional driving world models, conditioning signals such as high-definition (HD) maps, 3D bounding boxes, and reference frames are provided as transient per-frame or per-chunk inputs. While transient conditioning constrains immediate frame synthesis, it fails to establish a persistent memory state across extended rollouts. Recent retrieval-augmented models and object-centric attention mechanisms demonstrate the value of persistent scene context and isolated participant conditioning (Wu et al., 2025b; Xiao et al., 2025; Li et al., 2025; Jia et al., 2025; Seo et al., 2026; Yang & Zhang, 2026). However, existing methods deploy these mechanisms in isolation, failing to jointly

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.

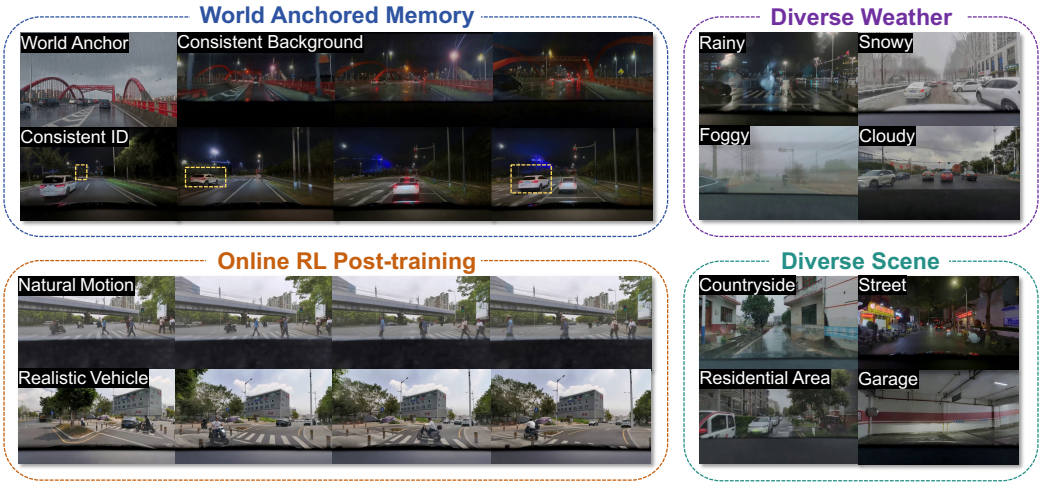

Figure 1: Overview of core capabilities in WAM-Sim. World anchored memory (top-left) anchors long-horizon rollouts, maintaining background consistency and persistent instance identities (e.g., tracked pedestrians, motor vehicles, and non-motorized vehicles) across extended horizons. Online RL Post-training (bottom-left) restores visual fidelity and enforces physical commonsense, synthesizing natural agent dynamics (e.g., pedestrian gait) and realistic vehicle motion. Furthermore, WAM-Sim generalizes across diverse weather conditions and environmental scenes (right).

preserve global route geometry, geolocated scene appearance, and individual object identities under dynamic mid-rollout interventions. Consequently, once generation extends beyond the initial context window, persistent states must be reconstructed from noisy self-generated histories, causing errors to compound over time. Overcoming this bottleneck requires a unified memory architecture that represents road structure, environmental appearance, and participant identity as distinct visual tokens, maintains each according to its temporal persistence, and retrieves each through mechanisms matched to its spatio-temporal validity domain.

To address this challenge, we present a driving world model anchored by a persistent world anchored memory. The memory organizes the driving world across three temporal and semantic scales: static structural tokens encoded from HD lane maps serve as persistent geometric anchors; read-only scene tokens retrieved from geolocated databases supply background appearance via temporally routed cross-attention; and instance tokens registered in a write-once appearance bank preserve individual object identities across chunk boundaries and occlusions through instance-masked cross-attention. Although this explicit memory substrate effectively eliminates structural and identity drift, injecting such dense, multi-scale conditioning signals inevitably shifts the generator off its pretrained distribution, compromising native visual sharpness and multi-agent physical plausibility. To overcome this trade-off, we introduce Drive DiffusionNFT, an online post-training framework that optimizes self-generated rollouts via flow-matching reinforcement learning guided by hierarchy-aligned visual rewards and a physical judge. Crucially, we uncover a practical synergy between explicit memory persistence and online policy optimization: the invariant memory manifold empirically confines the policy optimization space, helping mitigate semantic collapse and reward hacking during recursive rollouts, while online RL systematically restores fine-grained visual fidelity and enforces physical commonsense without compromising global stability.

Extensive evaluations on the standard nuScenes benchmark and a self-collected challenging long-horizon dataset demonstrate the effectiveness of WAM-Sim. On nuScenes, WAM-Sim achieves superior generation fidelity (FID $= 7.07$, FVD $= 42.01$) alongside strong downstream perception performance (19.67 mAP, 68.08 road mIoU). On minute-scale rollouts of self-collected data, it effectively mitigates quality–horizon decay, preserving high scene similarity (0.910) and instance identity coherence (0.681) across extended temporal gaps. Furthermore, comprehensive human preference studies and physical evaluations confirm that tightly coupling persistent hierarchical memory with online flow-matching post-training substantially mitigates structural and identity drift, synthesizing realistic and physically sound driving simulations.

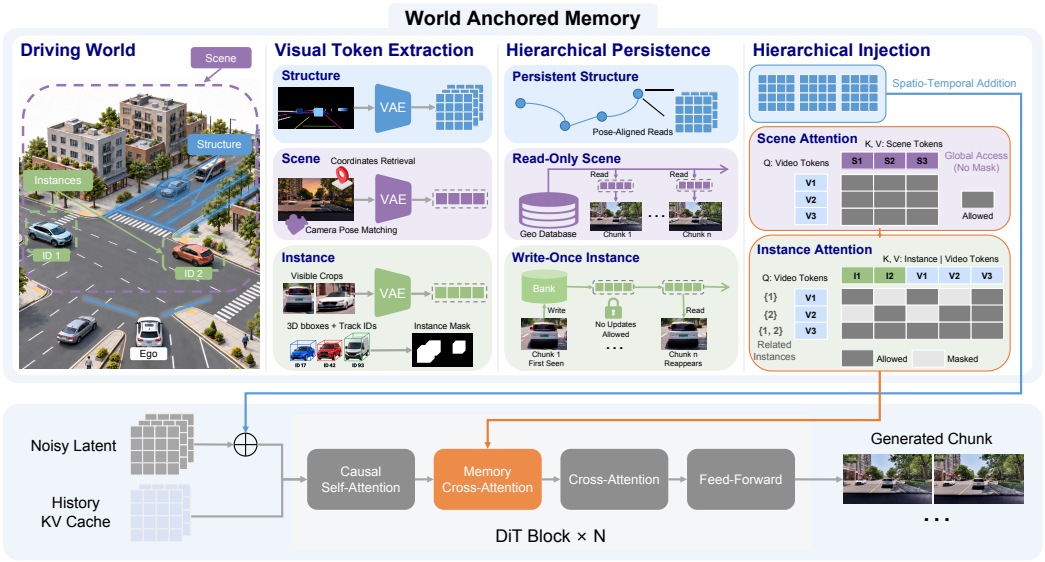

Figure 2: Overview of the WAM-Sim architecture and world anchored memory. Persistent context is structured across three temporal and semantic scales: 1) Structure: pose-aligned HD-map tokens injected via spatio-temporal addition to anchor road geometry; 2) Scene: geolocated background tokens retrieved from a database and accessed via temporally routed scene cross-attention; and 3) Instance: object appearance crops stored in a write-once appearance bank and retrieved via instance-masked cross-attention to isolate participant identities. The hierarchical memory state conditions a 3D Diffusion Transformer (DiT) operating on noisy latents and history KV caches to synthesize autoregressive, consistent driving video chunks.

2 Related Work

Driving World Models. Driving video generation has evolved from action- and trajectory-conditioned models (Hu et al., 2023; Russell et al., 2025; Gao et al., 2024c) to layout- and map-conditioned video synthesis (Wang et al., 2024a; Wen et al., 2024; Gao et al., 2024b;a; Zhao et al., 2025; Wang et al., 2024b; Li et al., 2024; Mei et al., 2024), autoregressive long-horizon generation (Jia et al., 2023; Hu et al., 2024; Zhang et al., 2025; Guo et al., 2024), geometry-aware 4D modeling (Yang et al., 2025; Jiang et al., 2025; Guo et al., 2025; Chen et al., 2024b), and driving-specialized foundation models (NVIDIA, 2025a; Ren et al., 2025; Ni et al., 2025; Basant et al., 2026). InstaDrive (Yang et al., 2026) propagates instance features to preserve identity, yet treats such evidence as transient per-chunk conditioning. Across these methods, maps, boxes, and retrieved context remain transient, with no persistent, addressable state binding geographic location and instance identities; we replace transient conditioning with a World Anchored Memory.

Memory and Long-Horizon Consistency. Persistence in interactive world models is realized through attention caches (Valevski et al., 2025; Decart & Etched, 2024; Parker-Holder et al., 2024; Google DeepMind, 2025), retrieval over past observations (Xiao et al., 2025; Yu et al., 2025; Li et al., 2025; Zhao et al., 2026; Wu et al., 2025b), or geographic grounding (Deng et al., 2024; Wang et al., 2025; Jia et al., 2025); orthogonal efforts extend context (Qiu et al., 2024; Kim et al., 2024; Henschel et al., 2025; Zhang & Agrawala, 2025), compress memory (Huang et al., 2025a; Oshima et al., 2025), mitigate exposure bias (Chen et al., 2024a; Yin et al., 2025; Huang et al., 2025b; Liu et al., 2026), or preserve identity (Yang & Zhang, 2026; Shen et al., 2026). Closest to ours, ConsisDrive (Yang & Zhang, 2026) adopts instance-masked attention, whereas we maintain a write-once appearance bank supporting occlusion re-retrieval; Spatial Retrieval (Jia et al., 2025) grounds generation geographically, whereas our scene memory is read-only and temporally routed. These mechanisms each address persistence in isolation; we unify geographic anchoring, cross-chunk scene state, and occlusion-surviving identity within a single memory substrate.

Reinforcement Learning for Visual Generation. Reward alignment of visual diffusion models has advanced from policy gradients and preference optimization (Black et al., 2023; Fan et al.,

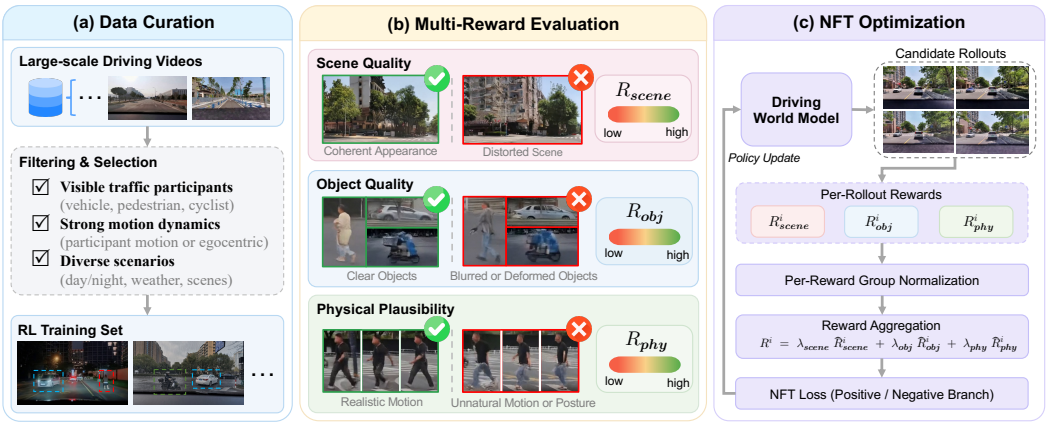

Figure 3: Overview of the Drive DiffusionNFT pipeline. (a) Data curation constructs a diverse RL training set with visible traffic participants and strong motion dynamics. (b) Multi-reward evaluation assesses scene quality, object quality, and physical plausibility. (c) NFT optimization normalizes and aggregates the rewards to update the driving world model.

2023; Wallace et al., 2023; Xu et al., 2023) to online group-relative flow-matching post-training (Liu et al., 2025; Xue et al., 2025; Zheng et al., 2025). Recently, RL post-training has been applied to video world models: RLVR-World (Wu et al., 2025a) leverages verifiable rewards for video generation, Reward Forcing (Lu et al., 2025) optimizes streaming autoregressive diffusion rollouts. Different from these approaches that focus on generic video rewards, WAM-Sim is the first to combine forward-process DiffusionNFT optimization with hierarchy-aligned driving rewards—unifying multi-scale visual fidelity with VLM-guided physical plausibility—to constrain policy search and systematically eliminate structural and quality decay on self-generated multi-chunk rollouts.

3 Method

As shown in Figure 2, we present a long-horizon driving world model (Section 3.1) built on two mutually reinforcing components: a persistent world anchored memory that anchors generation to a real geographic location and to per-instance identities (Section 3.2), and Drive DiffusionNFT, an online post-training stage that optimizes visual and physical realism on self-generated rollouts (Section 3.3). The two components are synergistic by design: the memory state delimits the manifold within which reward optimization operates, while the rewards are aligned with the memory hierarchy, so that the policy improves precisely the properties that memory persists.

3.1 Problem Formulation

Given a local HD-map e_c, 3D bounding-box agent controls a_c, a global ego pose P_c, and a persistent memory state $\mathcal{M}_c$, the model autoregressively synthesizes driving video chunks $\hat{I}_c \in \mathbb{R}^{C\times H\times W\times 3}$ across C temporal frames:

$$\hat{I}_c \sim p_\theta\Big(\cdot \,\Big|\, \hat{I}^{\text{tail}}_{c-1},\, e_c,\, a_c,\, G_c,\, \mathcal{M}_c\Big), \qquad \mathcal{M}_{c+1} = \text{Update}(\mathcal{M}_c,\, e_c, a_c, G_c, \hat{I}_c). \tag{1}$$

Here, a_c encodes 3D bounding-box layout sequences for surrounding agents, and G_c provides full global pose trajectories to background retrieval. Specifically, $\text{Update}(\cdot)$ is an append-only operation: it registers newly emerging traffic instances into the appearance bank upon initial observation and advances temporal segment indices, guaranteeing that existing memory entries are never overwritten or corrupted by self-generated video content. Long-horizon generation satisfies three requirements: 1) no *structural drift*—the video remains anchored to the road geometry and background appearance; 2) no *identity drift*—each tracked participant preserves its appearance across chunks, occlusion, and re-emergence; 3) no *quality decay*—visual and physical fidelity do not degrade with horizon.

3.2 WORLD ANCHORED MEMORY

An autonomous-driving world retains information at multiple scales: road topology, geolocated scene context, and traffic-participant identities. Because these states differ in semantics, spatial scope, and temporal persistence, we formalize the world state as a hierarchical memory tuple $\mathcal{M} = \left(\mathcal{M}^{\text{struct}}, \mathcal{M}^{\text{scene}}, \mathcal{M}^{\text{inst}}\right)$. Each memory tier operates under a dedicated persistence mechanism and is accessed via a specialized retrieval operator strictly aligned with its spatio-temporal domain of validity.

Memory Token Extraction. We extract memory tokens across three distinct semantic and temporal scales. First, structural tokens capture static road topology by projecting the per-frame HD-map sequence into the latent video grid via a dedicated map embedding layer. Second, scene tokens represent persistent background appearance; we partition the rollout into N_b temporal segments and retrieve a geolocated reference image using global pose P_c. Each reference image is VAE-encoded into L patch tokens $S_n \in \mathbb{R}^{L\times D}$, forming the unified scene memory matrix $S = \text{Concat}_{n=1}^{N_b} S_n \in \mathbb{R}^{N_b L\times D}$. Third, instance tokens encode object identity. For each tracked instance, After filtering observations by the bounding-box area range $[\rho_{\min}, \rho_{\max}]$, we divide the remaining frames into at most K_f temporal segments and select the largest observation from each. The selected crops are VAE-encoded and projected into $G_i \in \mathbb{R}^{K\times D}$, where $K = K_f K_p$ and K_p is the number of tokens per crop. A validity vector ν_i masks padded slots.

Hierarchical Persistence. The three memory tiers operate under persistence mechanisms strictly aligned with their temporal dynamics. Structural memory is trajectory-aligned: the global road topology remains invariant, while its per-frame HD-map representation dynamically transforms relative to the ego pose. Scene memory is geographically grounded and read-only: geolocated references in $\mathcal{M}^{\text{scene}}$ remain static throughout generation and cannot be overwritten by self-generated content, thereby anchoring environmental textures, architecture, and road surfaces to real-world coordinates. Instance memory follows a write-once policy: upon an agent's initial appearance, its visual tokens G_i are cached in a track-keyed bank $\mathcal{M}^{\text{inst}} = \{\tau_i \mapsto G_i\}$ indexed by non-parametric track identifiers τ_i. Locking G_i against subsequent overwrites prevents autoregressive error accumulation from corrupting object identities, guaranteeing seamless re-identification across extended occlusions and chunk boundaries.

Hierarchical Memory Injection. Rather than exposing all memory tokens to every video token, our hierarchical routing progressively conditions a shared video representation with geometry-, environment-, and identity-level information at their corresponding validity domains.

Structural memory provides trajectory-aligned geometric information. Since the HD-map representation is spatially aligned with the noisy video latent, it is incorporated as an additive condition, $X^{(0)} = Z_t + \mathcal{M}_{\text{struct}}$, where Z_t denotes the noisy video latent at flow timestep t. This establishes the geometry-aware video representation on which the subsequent memory routing operates. Scene memory operates at the temporal-segment level. For video token j, let $s(j)$ denote its corresponding segment. The routing mask is defined as $R^{\text{scene}}_{j,(n-1)P+r} = 0$ when $s(j) = n$ and otherwise $-\infty$, ensuring that each token retrieves scene information only from its associated temporal segment. Scene-memory retrieval is performed on the current geometry-conditioned video representation, producing an updated representation that incorporates the corresponding environmental context. Instance memory further specializes this shared representation within object-specific spatio-temporal regions. Given N tracked instances, we rasterize their bounding boxes onto the video-token grid to form an instance-membership matrix $\Pi \in \{0,1\}^{L\times N}$, where $\Pi_{j,i} = 1$ if and only if token j belongs to instance i. Only foreground tokens covered by tracked instances participate in the instance-memory branch, while background tokens bypass it. Each foreground token interacts only with video tokens belonging to the same instance and retrieves appearance information only from the corresponding entry in the instance memory bank. Cross-instance connections and invalid memory tokens are masked out. Importantly, instance-memory retrieval operates on the video representation already conditioned by both structural and scene information.

The resulting formulation constitutes a coarse-to-fine hierarchical conditioning process over a shared video representation: structural memory establishes trajectory-aligned geometry, scene memory introduces segment-specific environmental context, and instance memory preserves object-specific appearance within tracked regions. Although their validity domains differ, the three memories jointly form a single hierarchical memory system rather than independent parallel branches.

3.3 DRIVE DIFFUSIONNFT: SYNERGISTIC ONLINE POST-TRAINING

Although explicit memory persistence eliminates structural and identity drift, incorporating multi-scale conditioning introduces distribution shifts that can degrade visual fidelity and physical consistency. To resolve this trade-off, we introduce Drive DiffusionNFT, an online post-training framework optimizing self-generated rollouts via flow-matching reinforcement learning.

Hierarchical Reward Formulation. As shown in Figure 3 (b), our rewards mirror the memory hierarchy and the rollout reward R^i for candidate i integrates three multi-level criteria:

$$R^i = \lambda_{\text{scene}}\tilde{R}^i_{\text{scene}} + \lambda_{\text{obj}}\tilde{R}^i_{\text{obj}} + \lambda_{\text{phy}}\tilde{R}^i_{\text{phy}}, \tag{2}$$

where $\tilde{R}^i_m$ denotes the group-normalized score of reward component m. Specifically, scene-level perceptual quality averages the VBench Imaging Quality scores over sampled frames (Huang et al., 2024), and $R^i_{\text{scene}} = \frac{1}{|\mathcal{S}^i|}\sum_f q_{\text{scene}}(\hat{\mathbf{V}}^i_f)$. Object-level recognizability is evaluated on valid agent crops using binary classifiers trained on driving scenes from WorldLens (Liang et al., 2025), including the classes of vehicles and pedestrians. To avoid saturation of sigmoid confidence and retain greater score discrimination, we use normalized pre-sigmoid classifier logits, $R^i_{\text{obj}} = \frac{1}{N^i_{\text{obj}}}\sum_k \frac{1}{|\mathcal{C}^i_k|}\sum_{o\in\mathcal{C}^i_k}\tilde{\ell}_{\text{cls}}(o)$, where $\mathcal{C}^i_k$ denotes the valid crops of object k, and $\tilde{\ell}_{\text{cls}}(o)$ is the normalized logit from the corresponding class-specific classifier. Crops below a visible-area threshold are excluded. To explicitly enhance the physical realism of generated driving videos, we apply SoliReward (Lian et al., 2026) to access the general physical plausibility and driving scenario subject deformity, and thereby capturing physically implausible interactions and dynamics in driving scenes, as well as structural or temporal artifacts of traffic participants. For each generated video, the physical reward is given directly by its scalar reward score, $R^i_{\text{phy}} = s_{\text{Soli}}(\hat{\mathbf{V}}^i)$.

Flow-Matching Policy Optimization. Following DiffusionNFT (Zheng et al., 2025), we perform reward-guided online optimization directly over self-generated rollouts. For each condition $\mathbf{c}$, a group of G candidate rollouts is sampled from the collection policy π^{old} with the corresponding clean latents denoted by $\mathbf{x}_0$. For a noisy state $\mathbf{x}_t = \alpha_t\mathbf{x}_0 + \sigma_t\boldsymbol{\epsilon}$ with target velocity $\mathbf{v} = \dot{\alpha}_t\mathbf{x}_0 + \dot{\sigma}_t\boldsymbol{\epsilon}$ ($\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})$), the flow-matching objective updates policy velocity field $\mathbf{v}_\theta$ via:

$$\mathcal{L}_{\text{NFT}}(\theta) = \mathbb{E}_{\mathbf{c},\,\mathbf{x}_0\sim\pi^{\text{old}},\,t}\left[r\left\|\mathbf{v}^+_\theta - \mathbf{v}\right\|_2^2 + (1-r)\left\|\mathbf{v}^-_\theta - \mathbf{v}\right\|_2^2\right], \tag{3}$$

where $\mathbf{v}^+_\theta$ and $\mathbf{v}^-_\theta$ represent positive and negative implicit velocity predictions, and optimality weight $r = \sigma(R^i)$ converts group-relative rollout rewards into relative gradient updates.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUPS

Implementation Details. Our generator is a 3D DiT initialized from pretrained OmniDreams weights and trained in two stages: World-Anchored Memory pre-training and Drive DiffusionNFT post-training. Stage I sequentially optimizes trajectory-aligned structural conditioning, scene memory, and instance memory for 130k, 15k, and 5k steps, respectively. All use batch size 8 and context parallelism degree 8, with learning rates of 1×10^{-5}, 3×10^{-5}, and 3×10^{-4}. Starting from the full Stage I checkpoint, Stage II applies online flow-matching RL to self-generated rollouts using $G = 16$ and LoRA adapters (rank 256, LR 1×10^{-5}). We use AdamW throughout and train all experiments on 32 Ascend 910B NPUs.

Evaluation Metrics. We evaluate generation across four core dimensions: 1) visual fidelity via FID, FVD, and VBench Quality Score; 2) long-horizon consistency via Subject Coherence (instance identity) and Scene Similarity (background stability); 3) object & physical realism via detection-gated Subject Fidelity and VLM-guided Physical Plausibility; and 4) downstream perception on nuScenes via 3D detection (mAP, NDS) and BEV segmentation mIoU (road, vehicle).

4.2 MAIN BENCHMARK COMPARISONS

nuScenes Benchmark. Table 1 evaluates multi-view generation fidelity and downstream perception consistency on nuScenes. Lacking geolocated reference databases, WAM-Sim operates without

Table 1: Comparisons of controllable driving scene generation on the nuScenes dataset, grouped by generation paradigm. FID is computed over keyframes corresponding to all nuScenes validation.

Method	Synthesis Resolution	Fidelity		3D Detection		BEV Seg mIoU	
		FVD ↓	FID ↓	mAP ↑	NDS ↑	Road ↑	Vehicle ↑
Ori nuScenes	900 × 1600	-	-	35.54	41.21	73.66	34.82
(I) Image Generation Methods							
BEVControl	224 × 400	-	24.85	-	-	60.80	26.80
PerLDiff	256 × 384	-	13.36	15.24	24.05	61.26	27.13
DrivePTS	224 × 480	-	11.45	15.37	25.49	63.95	27.82
MagicDrive	224 × 400	-	16.20	12.30	23.32	61.05	27.01
(II) Video Generation Methods							
DriveDreamer	256 × 448	452.0	26.80	-	-	-	-
DriveDreamer-2	256 × 448	105.10	25.00	-	-	-	-
DrivingDiffusion	512 × 512	332.0	15.83	-	**33.10**	63.20	31.60
Panacea	256 × 512	139.0	16.96	11.58	22.31	55.78	22.74
DreamForge	224 × 400	224.8	14.61	13.01	22.16	65.27	28.36
MagicDrive-V2	848 × 1600	94.84	20.91	17.65	-	59.79	**32.73**
Ours	704 × 1280	**42.01**	**7.07**	**19.67**	29.95	**68.08**	31.53

Table 2: Evaluation on the Self-Collected Dataset. All methods generate videos conditioned on ego trajectories and agent bounding boxes.

Method	Visual Fidelity		VBench Metrics			Persistence		Object & Physics	
	FID ↓	FVD ↓	Quality ↑	Subject ↑	Background ↑	Sub. Coh. ↑	Scene Sim. ↑	Sub. Fid. ↑	Phys. Plaus. ↑
(I) Inference Only									
UniMLVG	44.21	753.16	75.75	83.43	89.77	0.591	0.708	39.19	46.57
OpenDWM	37.69	535.42	76.54	82.36	90.11	0.532	0.731	37.24	47.44
HorizonDrive	24.86	218.51	78.41	84.55	88.09	0.553	0.804	42.98	51.19
(II) SFT on Self-Collected Data									
OmniDreams	11.58	81.64	77.63	86.17	92.35	0.646	0.863	37.64	47.99
Ours	**10.95**	**74.20**	**78.71**	**86.98**	**92.54**	**0.681**	**0.910**	**44.00**	**55.96**

Table 3: World memory component ablation on the Self-Collected Dataset. Scene and Instance Memory improve generation consistency.

Scene	Instance	Sub. Coh. ↑	Scene Sim. ↑	FVD ↓
–	–	0.646	0.863	81.64
√	–	0.659	**0.911**	76.68
–	√	0.680	0.874	85.55
√	√	**0.681**	0.910	**74.20**

Table 4: Robustness to scene-memory dropout. Each reference is independently dropped at inference with probability p.

Drop Prob. p	Scene Sim. ↑	FVD ↓
0.0	**0.910**	**74.20**
0.1	0.908	77.03
0.3	0.901	78.60

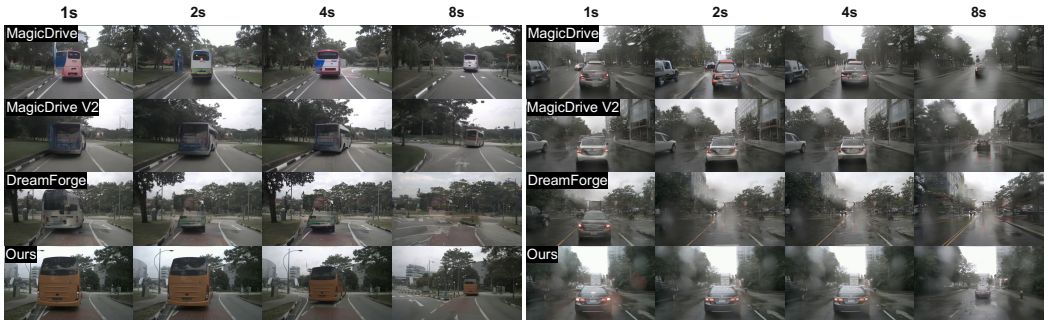

Figure 4: Qualitative comparison results on nuScenes video generation.

scene memory in this setting. At 704×1280 synthesis resolution, WAM-Sim achieves a visual fidelity (FID = 7.07, FVD = 42.01), outperforming the representative image and video generation paradigms. Furthermore, this high-fidelity synthesis translates into robust geometric control, yielding leading downstream 3D detection (19.67 mAP) and BEV road segmentation (68.08 mIoU).

Self-Collected Dataset. As shown in Table 2 and Figure 6, by anchoring rollouts with persistent world memory, WAM-Sim effectively mitigates quality–horizon decay, outperforming other base-

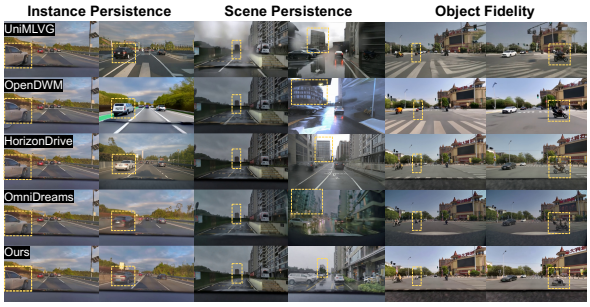

Figure 5: Qualitative comparison results on the self-collected dataset.



Figure 6: Diverse generation results on our self-collected dataset.

Table 5: Comparison of instance identity preservation on the test set.

Method	Long-range Coherence			Re-emergence				Avg.↑
	SC@1S↑	SC@2S↑	SC@3S↑	Sim.↑	Pass Rate↑	Occ-ID@1S↑	Occ-ID@2S↑	
Baseline (SFT)	0.605	0.579	0.546	0.780	0.461	0.781	0.743	0.642
ConsisDrive	0.634	0.604	0.597	0.805	0.510	0.799	0.812	0.680
Ours	**0.653**	**0.620**	**0.613**	**0.810**	**0.535**	**0.804**	**0.833**	**0.695**

Table 6: Ablation of reward components in Drive DiffusionNFT post-training.

R_{scene}	R_{obj}	R_{phy}	**Img.** (R_{scene}) ↑	**Classi.** (R_{obj}) ↑	**Phys.** (R_{phy}) ↑	**Sub. Fid.** ↑	**Phys. Plaus.** ↑
–	–	–	59.55	-0.012	3.770	41.74	50.91
√	–	–	62.68	0.064	3.727	42.25	51.13
√	√	–	61.59	0.273	3.789	43.97	53.54
√	√	√	61.48	0.271	4.052	44.00	55.96

lines with 10.95 FID and 74.20 FVD while securing superior background stability (0.910 Scene Similarity) and identity retention (0.681 Subject Coherence). Furthermore, Drive DiffusionNFT online post-training successfully enforces physical commonsense and object realism, yielding peak VBench Quality (78.71), Subject Fidelity (44.00), and Physical Plausibility (55.96).

4.3 Ablation and Discussion

Architecture Component Ablation. Table 3 disentangles the complementary roles of Scene Memory and Instance Memory on the self-collected dataset. Individually, Scene Memory anchors global environmental context, boosting Scene Similarity from 0.863 to 0.911 and reducing FVD from 81.64 to 76.68 over the memory-free baseline. Conversely, Instance Memory strengthens fine-grained identity persistence, elevating Subject Coherence to 0.680, though its standalone FVD degrades to 85.55 due to a lack of global structural bounds. Jointly integrating both tiers achieves optimal performance—yielding the lowest FVD, peak Subject Coherence, and near-peak Scene Similarity.

Scene Memory. We evaluate the robustness and cross-domain generalization of scene memory. Quantitatively, introducing reference dropout at inference demonstrates both active reliance and strong fault tolerance (Table 4); even at $p = 0.3$, Scene Similarity decays marginally from 0.910 to 0.901. Qualitatively, Figure 7(left) demonstrates that scene memory effectively disentangles persistent spatial layouts from transient environmental conditions: when conditioned on a dusk reference image ("Background"), WAM-Sim faithfully preserves the underlying architectural geometry and road layout throughout target daytime rollouts ("Ours"). Conversely, omitting scene memory ("w/o") leads to severe structural drift and hallucinated background geometry.

Instance Memory. We evaluate the capacity of write-once instance memory to maintain participant identity across extended horizons and occlusion–re-emergence events (Appendix E.3). We benchmark against the SFT baseline and ConsisDrive, which incorporates instance-level masking without a persistent memory state. In Table 5, while ConsisDrive increases average identity preservation from 0.642 to 0.680 over SFT via spatial masking alone, WAM-Sim surpasses both methods across all metrics. Because WAM-Sim and ConsisDrive share the same masking backbone, this performance gain isolates the effect of our write-once appearance bank, demonstrating that persistent memory prevents autoregressive identity drift across chunk boundaries. Qualitatively (Fig. 7, right), when an agent re-emerges post-occlusion, SFT suffers appearance mutation, whereas WAM-Sim recovers its original visual identity by retrieving stored track-keyed features.

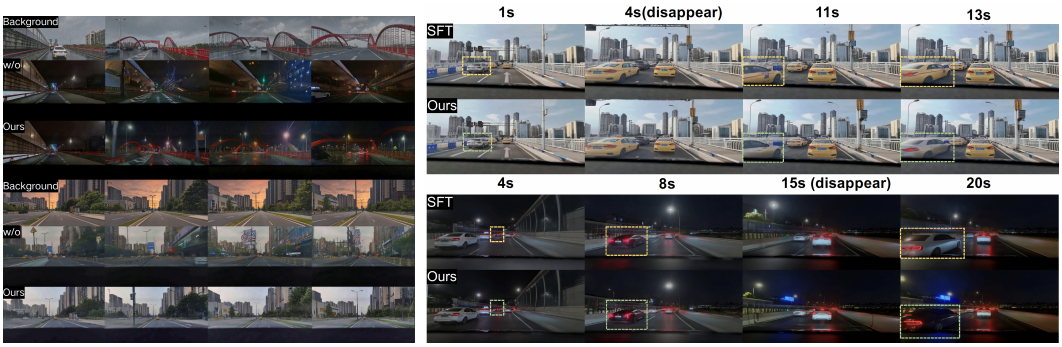

Figure 7: Qualitative comparison of scene- and instance-level persistence. WAM-Sim better preserves background appearance and object identity over long rollouts.

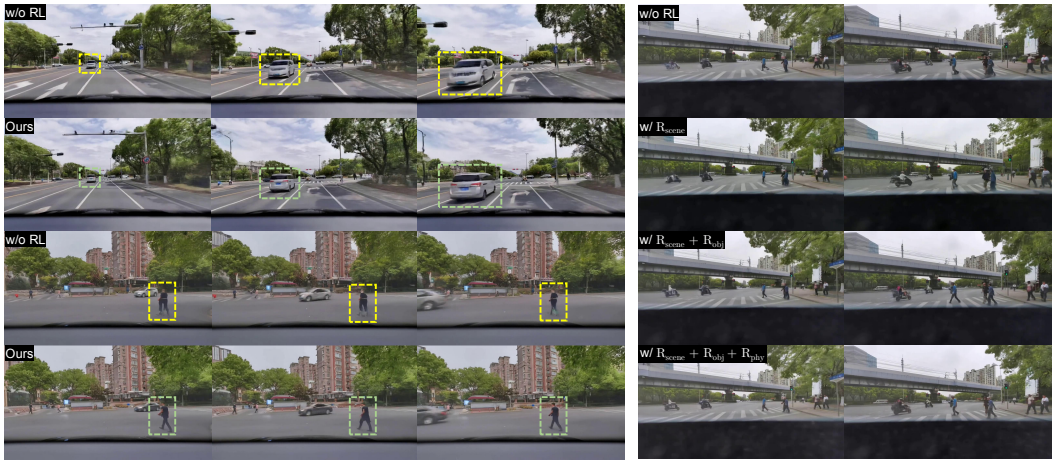

Figure 8: Qualitative ablation of online RL. Left: comparison between the world-memory model without RL and the final model. Right: cumulative reward ablation on the same scenario, progressively adding R_{scene}, R_{obj}, and R_{phy}.

Drive DiffusionNFT. Table 6 validates the hierarchical reward design in Drive DiffusionNFT. Incrementally incorporating R_{scene}, R_{obj}, and R_{phy} systematically boosts all optimization signals—improving VBench Imaging Quality ($59.55 \rightarrow 62.68$), Object Classification score ($-0.012 \rightarrow 0.271$), and SoliReward Physics score ($3.770 \rightarrow 4.052$). Crucially, these gains directly transfer to held-out evaluation metrics, elevating Subject Fidelity ($41.74 \rightarrow 44.00$) and Physical Plausibility ($50.91 \rightarrow 55.96$). This shows that multi-level reward alignment enhances physical realism and instance fidelity without collapsing into reward hacking. Figure 8 (right) qualitatively demonstrates the cumulative visual and physical gains across each reward tier.

Synergy Between Memory Bounds and Online RL. As shown in Figure 8 (left), explicit memory persistence and online policy optimization are mutually reinforcing: world anchored memory bounds the policy optimization manifold to prevent semantic drift across long rollouts, while Drive DiffusionNFT refines fine-grained visual sharpness and enforces physically plausible agent motions within this persistent world scaffold.

5 Conclusion

We presented WAM-Sim, a long-horizon driving world model that systematically mitigates structural drift, identity mutation, and quality decay. By unifying a persistent world anchored memory—anchoring road topology, geolocated appearance, and instance identity—with Drive DiffusionNFT online post-training, WAM-Sim establishes a persistent world state while enforcing visual sharpness and physical plausibility. Benchmark evaluations on nuScenes and self-collected driving datasets demonstrate superior generation fidelity, robust downstream perception, and temporal stability over extended rollouts. This synergy between invariant memory manifolds and reward-guided policy optimization offers a promising paradigm for closed-loop simulation.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, LaTeX formatting, and code debugging. They were not used to generate experimental data or scientific conclusions. All claims, citations, implementations, and reported results were reviewed and verified by the authors, who take full responsibility for the content of this work.

ETHICS STATEMENT

We use nuScenes in accordance with its license and employ a self-collected driving dataset solely for research purposes. Any public release will include only data permitted for redistribution and will exclude or anonymize faces, license plates, sensitive location information, and other personally identifiable content. The human evaluation collects only perceptual judgments of generated videos and reports results in aggregate. Generated videos and automated evaluation scores are intended for research and should not be used for safety-critical decision-making.

REPRODUCIBILITY STATEMENT

To ensure full reproducibility, we release our complete training and evaluation code, model checkpoints, and dataset processing pipelines. Detailed architectural configurations, optimization hyperparameters, and hardware environments are thoroughly documented in Appendices C and D.

REFERENCES

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. VideoPhy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024.

Aarti Basant, Amlan Kar, Despoina Paschalidou, Fangyin Wei, Francesco Ferroni, Guillermo Garcia Cobo, Haithem Turki, Huan Ling, Jaewoo Seo, James Lucas, Jay Zhangjie Wu, Jialiang Wang, Jonathan Lorraine, Jun Gao, Kai He, Katarina Tothova, Kevin Xie, Michał Tyszkiewicz, Qi Wu, Riccardo de Lutio, Ruilong Li, Sanja Fidler, Seung Wook Kim, Tianchang Shen, Tianshi Cao, Tobias Pfaff, William Lew, Xindi Wu, Xuanchi Ren, Yifan Lu, Yuxuan Zhang, Zan Gojcic, and Zian Wang. NVIDIA OmniDreams: Real-time generative world model for closed-loop autonomous vehicle simulation. *arXiv preprint arXiv:2606.03159*, 2026.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

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

Guibin Chen, Dixuan Lin, Jiangping Yang, et al. SkyReels-V2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025.

Rui Chen, Zehuan Wu, Yichen Liu, Yuxin Guo, Jingcheng Ni, Haifeng Xia, and Siyu Xia. UniMLVG: Unified framework for multi-view long video generation with comprehensive control capabilities for autonomous driving. *arXiv preprint arXiv:2412.04842*, 2024b.

Ke Cheng, Hanqiao Ye, Lei Shi, Yahui Liu, Yunhan Shen, Jingtao Dong, Zhenke Wang, Wenxuan Ao, Weixiang Xu, Kaining Huang, and Shuhan Shen. M^4World: A multi-view multimodal driving world model for interactive object manipulation and minute-long streaming. *arXiv preprint arXiv:2607.14005*, 2026.

Decart and Etched. Oasis: A universe in a transformer, 2024. Project page.

Boyang Deng, Richard Tucker, Zhengqi Li, Leonidas Guibas, Noah Snavely, and Gordon Wetzstein. Streetscapes: Large-scale consistent street view generation using autoregressive video diffusion. In *ACM SIGGRAPH Conference Papers*, 2024. doi: 10.1145/3641519.3657513.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. DPOK: Reinforcement learning for fine-tuning text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

Ruiyuan Gao, Kai Chen, Bo Xiao, Lanqing Hong, Zhenguo Li, and Qiang Xu. MagicDriveDiT: High-resolution long video generation for autonomous driving with adaptive control. *arXiv preprint arXiv:2411.13807*, 2024a.

Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. MagicDrive: Street view generation with diverse 3D geometry control. In *International Conference on Learning Representations*, 2024b.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. In *Advances in Neural Information Processing Systems*, 2024c.

Google DeepMind. Genie 3: A new frontier for world models, 2025. Google DeepMind technical report.

Jiazhe Guo, Yikang Ding, Xiwu Chen, Shuo Chen, Bohan Li, Yingshuang Zou, Xiaoyang Lyu, Feiyang Tan, Xiaojuan Qi, Zhiheng Li, and Hao Zhao. DiST-4D: Disentangled spatiotemporal diffusion with metric depth for 4D driving scene generation. *arXiv preprint arXiv:2503.15208*, 2025.

Xi Guo, Chenjing Ding, Haoxuan Dou, Xin Zhang, Weixuan Tang, and Wei Wu. InfinityDrive: Breaking time limits in driving world models. *arXiv preprint arXiv:2412.01522*, 2024.

Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. StreamingT2V: Consistent, dynamic, and extendable long video generation from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. GAIA-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Xiaotao Hu, Wei Yin, Mingkai Jia, Junyuan Deng, Xiaoyang Guo, Qian Zhang, Xiaoxiao Long, and Ping Tan. DrivingWorld: Constructing world model for autonomous driving via video GPT. *arXiv preprint arXiv:2412.19505*, 2024.

Junchao Huang, Xinting Hu, Boyao Han, Shaoshuai Shi, Zhuotao Tian, Tianyu He, and Li Jiang. Memory forcing: Spatio-temporal memory for consistent scene generation on Minecraft. *arXiv preprint arXiv:2510.03198*, 2025a.

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*, 2025b.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Fan Jia, Weixin Mao, Yingfei Liu, Yucheng Zhao, Yuqing Wen, Chi Zhang, Xiangyu Zhang, and Tiancai Wang. ADriver-I: A general world model for autonomous driving. *arXiv preprint arXiv:2311.13549*, 2023.

Xiaosong Jia, Chenhe Zhang, Yule Jiang, Songbur Wong, et al. Spatial retrieval augmented autonomous driving. *arXiv preprint arXiv:2512.06865*, 2025.

Junpeng Jiang, Gangyi Hong, Miao Zhang, Hengtong Hu, Kun Zhan, Rui Shao, and Liqiang Nie. DiVE: Efficient multi-view driving scenes generation based on video diffusion transformer. *arXiv preprint arXiv:2504.19614*, 2025.

Wooyeol Kim, Junho Park, Janghyeon Lee, Wonjun Hwang, Hyunwoo J. Kim, and Kyungmin Jo. FIFO-diffusion: Generating infinite videos from text without training. In *Proceedings of the International Conference on Machine Learning*, 2024.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. VMem: Consistent interactive video scene generation with surfel-indexed view memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. DrivingDiffusion: Layout-guided multi-view driving scene video generation with latent diffusion model. In *European Conference on Computer Vision*, 2024.

Jiesong Lian, Ruizhe Zhong, Zixiang Zhou, Xiaoyue Mi, Yixue Hao, Yuan Zhou, Qinglin Lu, Long Hu, and Junchi Yan. SoliReward: Mitigating susceptibility to reward hacking and annotation noise in video generation reward models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12798–12807, 2026.

Aoyu Liang, Lingdong Kong, Tong Yan, Hao Liu, Yang Yang, et al. WorldLens: Full-spectrum evaluations of driving world models in real world. *arXiv preprint arXiv:2512.10958*, 2025.

Wang Lin, Liyu Jia, Wentao Hu, Kaihang Pan, Zhongqi Yue, Wei Zhao, Jingyuan Chen, Fei Wu, and Hanwang Zhang. Reasoning physical video generation with diffusion timestep tokens via reinforcement learning. *arXiv preprint arXiv:2504.15932*, 2025.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online RL. In *Advances in Neural Information Processing Systems*, 2025.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *International Conference on Learning Representations*, 2026.

Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, et al. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. *arXiv preprint arXiv:2512.04678*, 2025.

Jianbiao Mei, Tao Hu, Xuemeng Yang, Licheng Wen, Yu Yang, Tiantian Wei, Yukai Ma, Min Dou, Botian Shi, and Yong Liu. DreamForge: Motion-aware autoregressive video generation for multi-view driving scenes. *arXiv preprint arXiv:2409.04003*, 2024.

Jingcheng Ni, Yuxin Guo, Yichen Liu, Rui Chen, Lewei Lu, and Zehuan Wu. OpenDWM: Open driving world models, 2025. Open-source codebase, SenseTime.

NVIDIA. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025a.

NVIDIA. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025b.

Yuta Oshima, Yusuke Iwasawa, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. WorldPack: Compressed memory improves spatial consistency in video world modeling. *arXiv preprint arXiv:2512.02473*, 2025.

Jack Parker-Holder, Philip Ball, Jake Bruce, et al. Genie 2: A large-scale foundation world model, 2024. Google DeepMind blog post.

Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients. In *Proceedings of the International Conference on Machine Learning*, 2024.

Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. FreeNoise: Tuning-free longer video diffusion via noise rescheduling. In *International Conference on Learning Representations*, 2024.

Xuanchi Ren, Yifan Lu, Tianshi Cao, Ruiyuan Gao, Shengyu Huang, Amirmojtaba Sabour, Tianchang Shen, Tobias Pfaff, Jay Zhangjie Wu, Runjian Chen, Seung Wook Kim, Jun Gao, Laura Leal-Taixé, Mike Chen, Sanja Fidler, and Huan Ling. Cosmos-drive-dreams: Scalable synthetic driving data generation with world foundation models. *arXiv preprint arXiv:2506.09042*, 2025.

Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. GAIA-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*, 2025.

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

Liao Shen, Wentao Jiang, Yiran Zhu, Jiahe Li, Tiezheng Ge, Zhiguo Cao, and Bo Zheng. Identity-preserving image-to-video generation via reward-guided optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27281–27290, 2026.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *International Conference on Learning Representations*, 2025.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. *arXiv preprint arXiv:2311.12908*, 2023.

Jiawei Wang, Haowei Sun, Xintao Yan, Shuo Feng, Jun Gao, and Henry X. Liu. TeraSim-World: Worldwide safety-critical data synthesis for end-to-end autonomous driving. *arXiv preprint arXiv:2509.13164*, 2025.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. DriveDreamer: Towards real-world-driven world models for autonomous driving. In *European Conference on Computer Vision*, 2024a.

Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea: Panoramic and controllable video generation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Jialong Wu, Shaofeng Yin, Ningya Feng, and Mingsheng Long. RLvr-World: Training world models with reinforcement learning. *arXiv preprint arXiv:2505.13934*, 2025a.

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*, 2025b.

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

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. DanceGRPO: Unleashing GRPO on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Yu Yang, Alan Liang, Jianbiao Mei, Yukai Ma, Yong Liu, and Gim Hee Lee. X-Scene: Large-scale driving scene generation with high fidelity and flexible controllability. In *Advances in Neural Information Processing Systems*, 2025.

Zhuoran Yang and Yanyong Zhang. ConsisDrive: Identity-preserving driving world models for video generation by instance mask. *arXiv preprint arXiv:2602.03213*, 2026.

Zhuoran Yang, Xi Guo, Chenjing Ding, Chiyu Wang, Wei Wu, and Yanyong Zhang. InstaDrive: Instance-aware driving world models for realistic and consistent video generation. *arXiv preprint arXiv:2602.03242*, 2026.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

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 *SIGGRAPH Asia Conference Papers*, 2025.

Kaiwen Zhang, Zhenyu Tang, Xiaotao Hu, Xingang Pan, Xiaoyang Guo, Yuan Liu, Jingwei Huang, Li Yuan, Qian Zhang, Xiaoxiao Long, Xun Cao, and Wei Yin. Epona: Autoregressive diffusion world model for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Lvmin Zhang and Maneesh Agrawala. Packing input frame context in next-frame prediction models for video generation. In *Advances in Neural Information Processing Systems*, 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10412–10420, 2025.

Jinjing Zhao, Fangyun Wei, Zhening Liu, Hongyang Zhang, Chang Xu, and Yan Lu. Spatia: Video generation with updatable spatial memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4245–4257, 2026.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. DiffusionNFT: Online diffusion reinforcement with forward process. *arXiv preprint arXiv:2509.16117*, 2025.

Yang Zhou, Hao Shao, Letian Wang, Zhuofan Zong, Hongsheng Li, and Steven L. Waslander. DrivingGen: A comprehensive benchmark for generative video world models in autonomous driving. *arXiv preprint arXiv:2601.01528*, 2026.

A Extended Related Work

Driving World Models. Controllable driving video generation was pioneered by action-conditioned world models such as GAIA-1 and GAIA-2 (Hu et al., 2023; Russell et al., 2025), high-fidelity trajectory-conditioned prediction in Vista (Gao et al., 2024c), and multi-view generators conditioned on BEV layouts, HD maps, and 3D bounding boxes (Wang et al., 2024a; Wen et al., 2024; Gao et al., 2024b;a; Zhao et al., 2025). Autoregressive token-based approaches extend such models to longer horizons (Jia et al., 2023; Hu et al., 2024; Zhang et al., 2025), while geometry-aware methods improve spatial and 4D consistency (Yang et al., 2025; Jiang et al., 2025; Guo et al., 2025; Chen et al., 2024b). Two recent trends are most relevant to our setting. First, world foundation models specialized for driving—Cosmos and Cosmos-Drive-Dreams (NVIDIA, 2025a; Ren et al., 2025), and most recently OmniDreams (Basant et al., 2026)—achieve photorealistic, interactive, real-time rollouts under world-scenario control, with OmniDreams attaining minute-long closed-loop simulation through a streaming KV cache and self-forcing-style distillation. Second, a line of work explicitly targets long driving sequences: InfinityDrive introduces memory injection and retention for minute-scale driving video (Guo et al., 2024). Despite this progress, maps, boxes, and retrieved context are re-specified per chunk and therefore constrain only instantaneous generation; what persists across chunks—an addressable state binding a real location and per-instance identities—remains implicit (e.g., in attention caches) rather than explicit and controllable. We upgrade transient conditioning to a persistent world anchored memory while retaining minute-scale intervenability.

Memory and Long-Horizon Consistency. Interactive world models increasingly rely on explicit memory. Genie 3 exhibits emergent scene memory over minute-scale rollouts but, by its own account, cannot reproduce real-world locations with geographic accuracy (Google DeepMind, 2025). Retrieval-based memories index past observations for consistency: WorldMem maintains a retrievable bank indexed by camera pose (Xiao et al., 2025), VMem indexes past views by surfel geometry (Li et al., 2025), Spatia preserves an updatable 3D point cloud maintained with visual SLAM (Zhao et al., 2026), and video world models with long-term spatial memory support view re-identification over revisits (Wu et al., 2025b). Complementary to these generic mechanisms, geolocated generation grounds synthesis in real places: Streetscapes synthesizes street views of an on-the-fly city rather than actual locations (Deng et al., 2024); TeraSim-World grounds safety-critical data synthesis at arbitrary coordinates but channels real-world appearance through a VLM text bottleneck restricted to static elements (Wang et al., 2025); and, concurrently with our work, Spatial-RAG-AD retrieves geographic imagery along the ego route and injects it as persistent cross-attention context into driving video DiTs (Jia et al., 2025). On the training side, long-horizon methods either extend context windows (FreeNoise (Qiu et al., 2024), FIFO-Diffusion (Kim et al., 2024), StreamingT2V (Henschel et al., 2025)) or close the exposure-bias gap with self-generated rollouts, from Diffusion Forcing (Chen et al., 2024a) and CausVid (Yin et al., 2025) to Self-Forcing (Huang et al., 2025b) and Rolling Forcing (Liu et al., 2026); OmniDreams concurrently adopts self-forcing distillation for real-time driving rollouts (Basant et al., 2026). Instance-level consistency within single clips is pursued by identity-preserving driving generation (Yang & Zhang, 2026) and reference-image subject consistency methods (Shen et al., 2026). Existing works, however, provide either scene-level memory without real geographic anchors, geographic grounding without a persistent cross-chunk memory state, or instance consistency without occlusion-surviving, write-once identity evidence under counterfactual manipulation. We unify the three: a coordinate-indexed scene memory reused for the entire rollout together with a write-once appearance bank keyed by track identity, retrieved through banked instance-masked attention at chunk boundaries and upon re-emergence after occlusion.

Reinforcement Learning for Visual Generation. RL-based alignment of diffusion and flow models has progressed from policy-gradient fine-tuning of denoising MDPs (DDPO (Black et al., 2023); DPOK and Diffusion-DPO for image alignment (Fan et al., 2023; Wallace et al., 2023); VADER for video (Prabhudesai et al., 2024)), through video-scale preference optimization (SkyReels-V2 (Chen et al., 2025)), to online group-relative policy optimization for flow matching: Flow-GRPO introduced ODE-to-SDE exploration (Liu et al., 2025), DanceGRPO extended it to general visual generation (Xue et al., 2025), Phys-AR optimizes physical velocity-error rewards for video generation (Lin et al., 2025), and RLVR-World trains world models with RL (Wu et al., 2025a); reward-guided optimization has likewise been used for identity-preserving image-to-video generation (Shen et al., 2026). On the evaluation side, physical commonsense benchmarks such as VideoPhy (Bansal et al., 2024), PhysGame, and T2VPhysBench diagnose violations without training, while VLM

judges including Cosmos-Reason1 (NVIDIA, 2025b) and VBench-2.0 Motion Rationality establish that vision-language models can score physical plausibility; driving-specific suites such as DrivingGen (Zhou et al., 2026) and WorldLens (Liang et al., 2025) provide comprehensive but point-estimate evaluations on fixed short horizons. To our knowledge, we are the first to combine online flow-matching RL with driving-specific rewards—hierarchical visual fidelity at scene and instance levels together with a VLM-based physical plausibility judge—optimized directly on self-generated multi-chunk rollouts, and the first to evaluate long-horizon driving world models with quality–horizon decay curves, counterfactual ID-surgery, and mid-rollout intervention protocols.

B Additional Experiments

B.1 Human Evaluation

Experimental Setup and Protocol. We evaluate human perceptual preference on 60 representative driving scenarios sampled from the test set. Fifteen independent expert assessors with backgrounds in autonomous driving evaluate WAM-Sim against four baselines: UniMLVG, OpenDWM–CTSD3.5, HorizonDrive, and OmniDreams. To eliminate bias, all method identities are double-blinded, and video presentation order and spatial placement are randomized. Assessors perform two complementary tasks: (i) forced-choice pairwise preference voting (Win, Tie, or Loss) for side-by-side video pairs, and (ii) five-point Mean Opinion Score (MOS) rating across four core dimensions: *Visual Fidelity* (perceptual quality and temporal coherence), *Identity Consistency* (long-horizon participant appearance preservation across occlusions), *Physical Realism* (plausibility of agent dynamics and road interactions), and *Location Anchoring* (alignment with spatial and geographic context).

Score Aggregation and Inter-Rater Reliability. Pairwise preference judgments are aggregated per scenario via majority voting across the 15 assessors to determine scenario-level outcomes. MOS ratings are averaged per scenario across assessors and subsequently macro-averaged over all 60 scenarios. Inter-rater reliability across the 15 judges is validated using Fleiss' κ on categorical pairwise preferences, demonstrating strong agreement across all baseline comparisons: $\kappa_{\text{UniMLVG}} = 0.927$, $\kappa_{\text{OpenDWM}} = 0.864$, $\kappa_{\text{HorizonDrive}} = 0.939$, and $\kappa_{\text{OmniDreams}} = 0.972$.

Quantitative Results and Statistical Analysis. As summarized in Table 7, WAM-Sim substantially outperforms all baselines in pairwise preference, achieving scenario-level win rates of 98.3% against UniMLVG ($59/0/1$), 98.3% against OpenDWM–CTSD3.5 ($59/0/1$), 71.7% against HorizonDrive ($43/1/16$), and 86.7% against OmniDreams ($52/0/8$). Furthermore, WAM-Sim secures the highest MOS across all four perceptual dimensions (Fidelity: 4.30, Identity: 4.40, Physics: 4.50, Location: 4.50). To assess statistical significance, we conduct two-sided Wilcoxon signed-rank tests ($\alpha = 0.05$) on paired scenario-level MOS values (Table 8). WAM-Sim achieves statistically significant gains over UniMLVG, OpenDWM–CTSD3.5, and OmniDreams across all dimensions ($p < 10^{-9}$). Compared to HorizonDrive (the strongest baseline), WAM-Sim yields statistically significant improvements in Identity ($p = 7.65 \times 10^{-4}$), Physics ($p = 4.32 \times 10^{-8}$), and Location ($p = 3.41 \times 10^{-6}$), confirming that explicit world memory and online RL post-training primarily enhance long-horizon temporal consistency, physical plausibility, and spatial anchoring.

Table 7: Human evaluation on 60 driving scenarios assessed by 15 experts. Pairwise preference reports scenario-level win/tie/loss rates against WAM-Sim. MOS is reported on a 1–5 scale.

	Pairwise Preference (%)			**Mean Opinion Score**			
Method	**Win**	**Tie**	**Loss**	**Fidelity**	**Identity**	**Physics**	**Location**
UniMLVG	98.3	0.0	1.7	2.10	2.20	1.73	1.95
OpenDWM–CTSD3.5	98.3	0.0	1.7	1.80	1.80	1.38	1.87
HorizonDrive	71.7	1.7	26.7	4.10	4.00	4.00	3.90
OmniDreams	86.7	0.0	13.3	3.10	3.00	2.77	2.87
WAM-Sim (Ours)	–	–	–	**4.30**	**4.40**	**4.50**	**4.50**

Table 8: Statistical significance of scenario-level MOS differences. Two-sided Wilcoxon signed-rank tests compare WAM-Sim with each baseline using 60 paired scenario-level scores per dimension. Values below 0.05 meet the stated significance threshold.

Dimension	UniMLVG	OpenDWM–CTSD3.5	HorizonDrive	OmniDreams
Fidelity	7.29×10^{-12}	7.52×10^{-12}	0.0546	2.40×10^{-10}
Identity	7.58×10^{-12}	7.28×10^{-12}	7.65×10^{-4}	2.58×10^{-10}
Physics	7.52×10^{-12}	5.96×10^{-12}	4.32×10^{-8}	6.49×10^{-12}
Location	2.78×10^{-12}	6.25×10^{-12}	3.41×10^{-6}	5.17×10^{-12}

Table 9: Efficiency comparison of driving world models during long-horizon rollouts.

Method	Resolution	Params (B)	Latency (s) ↓	Throughput (FPS) ↑	VRAM (GB) ↓
Horizondrive	384 × 768	1.3	3.51	4.31	**20.11**
UniMLVG	256 × 448	2.0	10.70	1.54	25.32
OpenDWM-CTSD3.5	256 × 448	2.5	14.98	1.08	32.67
Omnidreams	704 × 1280	2.0	0.65	8.00	45.71
Ours	256 × 448	3.1	**0.60**	**18.63**	45.21
Ours	704 × 1280	3.1	1.19	6.94	47.45

B.2 Runtime and Overhead.

As detailed in Table 9, WAM-Sim maintains high computational efficiency across resolutions despite incorporating explicit hierarchical memory modules. At 704×1280, WAM-Sim sustains 6.94 FPS, retaining 86.8% of OmniDreams' throughput (8.00 FPS) with a negligible 3.8% increase in peak VRAM (47.45 GB vs. 45.71 GB). At 256×448, WAM-Sim reaches 18.63 FPS, outperforming same-resolution baselines UniMLVG (1.54 FPS) and OpenDWM–CTSD3.5 (1.08 FPS) by $12.1\times$ and $17.3\times$, respectively. Although structural, scene, and instance memory branches expand total parameter capacity from 2.0B to 3.1B (+55.0%), memory retrieval via temporally routed and instance-masked cross-attention incurs minimal latency (+0.54s per 93-frame chunk).

B.3 Quality-Horizon Decay Analysis.

We evaluate temporal stability across 10 s, 30 s, and 60 s autoregressive rollouts (Fig. 9). Without world memory, compounding autoregressive errors cause severe degradation from 10 s to 60 s: Subject Coherence drops from 0.535 to 0.322, Scene Similarity degrades from 0.791 to 0.697, and FVD deteriorates from 197.77 to 286.54. In contrast, WAM-Sim effectively bounds drift across extended horizons, maintaining high instance coherence (0.547), scene similarity (0.828), and visual fidelity (FVD = 137.92) at 60 s.

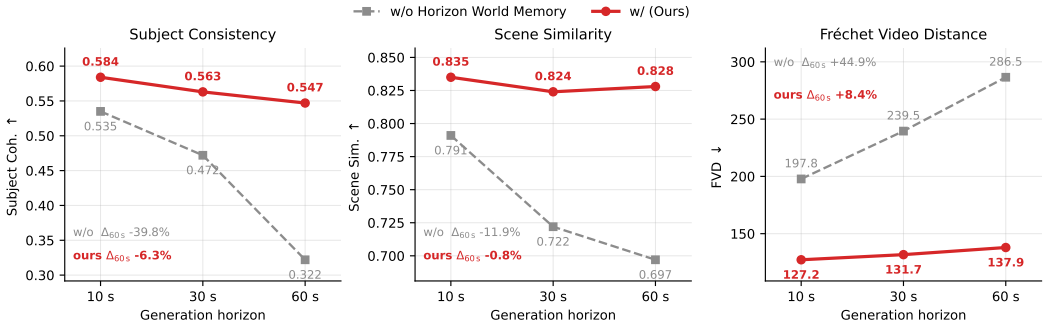

Figure 9: Quality versus generation horizon. World memory reduces quality degradation from 10 s to 60 s.

C Implementation Details

We present the architectural specifications, stage-wise training workflow, and reward formulations of WAM-Sim. The framework is trained in four progressive stages: structural control pre-training, scene-memory alignment, instance-memory registration, and online reinforcement post-training. Key model configurations and training parameters are summarized in Tables 11, 12, and 10.

C.1 Model Configuration

WAM-Sim is built on a 28-block 3D Diffusion Transformer (DiT) backbone with a hidden dimension of 2048 and 16 attention heads ($D_{\text{head}} = 128$), conditioned on text embeddings from a frozen Qwen2.5-VL-7B encoder. Input videos at 704×1280 resolution (93 frames at 10 FPS) are encoded by a Wan2.1 Video VAE with $8\times$ spatial and $4\times$ temporal compression into a $24 \times 88 \times 160$ latent grid. Applying $2 \times 2 \times 1$ spatio-temporal patchification yields 84,480 video tokens per chunk, modulated via 3D RoPE and timestep-conditioned adaLN-zero layers. As summarized in Table 11, Hierarchical World Memory conditions this shared sequence across three scale-specific tiers: (i) *Structural Memory* injects VAE-aligned HD map latents as additive conditioning to establish trajectory-aligned geometry; (ii) *Scene Memory* retrieves one geolocated reference frame (P tokens) per temporal segment, injecting background appearance via segment-masked cross-attention with zero-initialized adaLN gating; and (iii) *Instance Memory* extracts multi-frame appearance crops ($K_f = 4$ frames, $K_p = 64$ tokens per crop, capping at 32 active instances) into a write-once track-keyed bank, updating foreground tokens via instance-masked attention while bypassing background regions.

Table 10: Training settings for Drive DiffusionNFT.

Hyperparameter	**Value**
Group size G	16
Sampling steps	35
Guidance scale	1.0
Flow shift	5.0
NFT mixing coefficient β	1.0
Training timestep fraction	0.8
Reward weights $(\lambda_{\text{scene}}, \lambda_{\text{obj}}, \lambda_{\text{phy}})$	0.4, 0.3, 0.3
LoRA rank / alpha	256 / 256
Optimizer	AdamW
Learning rate	1×10^{-5}
Weight decay	1×10^{-2}

C.2 Training Configuration

WAM-Sim is trained via a progressive four-stage curriculum that incrementally introduces structural guidance, hierarchical memory routing, and online post-training without destabilizing the pretrained backbone. Stage I pre-trains structural control for 130k steps (LR $= 1 \times 10^{-5}$) using rectified flow to establish trajectory-aligned road geometry. Stage II aligns scene memory for 15k steps (LR $= 3 \times 10^{-5}$) to inject geolocated background representations via segment-masked cross-attention. Stage III optimizes instance memory for 5k steps (LR $= 3 \times 10^{-4}$) to register write-once appearance tokens into the track-keyed memory bank. Finally, Stage IV executes online post-training via Drive DiffusionNFT (LR $= 1 \times 10^{-5}$) with LoRA adapters to align self-generated rollouts against multi-scale visual and physical rewards. All stages utilize the AdamW optimizer with constant learning rates, $8\times$ context parallelism, and bfloat16 precision; complete optimization hyperparameters are summarized in Table 12.

C.3 Details of Drive DiffusionNFT Training

Drive DiffusionNFT aligns self-generated rollouts by optimizing a composite objective structured across three hierarchical visual and physical reward components. First, global visual quality R_{scene} is computed as the frame-averaged imaging quality score using a SPAQ-pretrained MUSIQ model

Table 11: Model configuration. Shared backbone architecture and configurations of the structural, scene, and instance memory components.

Configuration	Structural Memory	Scene Memory	Instance Memory
Shared Video Representation and Backbone			
Input resolution		704×1280	
Frames per window		93	
Frame rate		10 FPS	
Video VAE		Wan2.1 Video VAE	
VAE latent channels		16	
Spatial / temporal compression		$8\times$ / $4\times$	
Latent shape		$24 \times 88 \times 160$	
Spatial / temporal patch size		2×2 / 1	
Video-token grid		$24 \times 44 \times 80$	
Number of video tokens		84,480	
Hidden dimension		2048	
Number of DiT blocks		28	
Attention heads		16	
Head dimension		128	
MLP ratio / hidden dimension		4 / 8192	
Position encoding		Learnable 3D RoPE (t, h, w)	
Normalization		Timestep-conditioned adaLN-zero	
Text encoder		Frozen Qwen2.5-VL-7B (Cosmos-Reason1)	
Text embedding dimension		1024	
World Anchored Memory			
Memory scope	Trajectory-level	Temporal-segment-level	Instance-level
Condition source	HD map	Geolocated scene reference	Tracked object appearance
Validity domain	Video-token grid	Corresponding temporal segment	Corresponding instance region
Representation	VAE-aligned map latent	Scene-memory tokens	Appearance-memory tokens
Injection / retrieval	Additive conditioning	Segment-masked cross-attn	Instance-masked attention
Persistence	Trajectory aligned	Read-only after retrieval	Persistent across chunks
Reference frames	–	1 per segment	$K_f = 4$
Reference resolution	–	704×1280	128×128
Tokens per reference	–	P	$K_p = 64$
Tokens per memory entry	–	P	$K = 256$
Maximum entries	–	N_b segments	32 instances
Update strategy	Updated with trajectory	Fixed after retrieval	Write once at first appearance
Residual modulation	–	Zero-initialized adaLN gate	Zero-initialized adaLN gate

(Huang et al., 2024). Second, object recognizability R_{obj} evaluates dynamic traffic participants (pedestrians, cyclists, and vehicles). Subject crops extracted with $1.2\times$ bounding-box expansion are filtered by boundary margin, area, and occlusion, exposure-normalized against ground-truth frames to mitigate luminance gaming, and scored by a frozen ResNet-50 classifier (Liang et al., 2025) using dataset-calibrated z-normalized logits. Third, physical plausibility R_{phy} is evaluated via SoliReward (Lian et al., 2026), an InternVL3-1B model fine-tuned for motion realism and structural deformation assessment under its standard evaluation protocol.

The scalar rewards are group-normalized and linearly aggregated into the overall rollout objective. Optimization parameters, LoRA adapter configurations, and reward weighting factors $(\lambda_{\text{scene}}, \lambda_{\text{obj}}, \lambda_{\text{phy}})$ for online flow-matching post-training are detailed in Table 10.

Table 12: Training configuration. Stage-wise training settings for structural control, scene memory, instance memory, and reinforcement fine-tuning.

Training Setting	Stage I: Structural	Stage II: Scene	Stage III: Instance	Stage IV: Reinforcement
Training target	Structural control	Scene memory	Instance memory	Generation quality
Initialized from	Pretrained backbone	Stage I	Stage II	Stage III
Trainable modules	Self-Attn	Scene.Cross-attn	Instance.Cross-Attn	Self-Attn
Training objective	Rectified flow	Rectified flow	Rectified flow	DiffusionNFT loss
Learning rate	1×10^{-5}	3×10^{-5}	3×10^{-4}	1×10^{-5}
Optimizer	AdamW	AdamW	AdamW	AdamW
Weight decay	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-2}
LR schedule	constant	constant	constant	constant
Training steps	130k	15k	5k	300
Global batch size	4	4	4	-
Training resolution	704×1280	704×1280	704×1280	704×1280
Frames per sample	93	93	93	93
Latent frames	24	24	24	24
Context parallel size	8	8	8	8
Precision	bf16	bf16	bf16	bf16
Training hardware	910b	910b	910b	910b

D Dataset Construction

Training a long-horizon driving world model requires more than large-scale video data. The model must learn temporally consistent visual dynamics together with structured geometric conditions, persistent object identities, and semantic scene descriptions. We therefore construct a large-scale in-house autonomous-driving dataset and apply a multi-stage processing pipeline to transform raw driving logs into structured training samples. This section describes the data sources, construction pipeline, and final training and evaluation sets.

D.1 Data Source

Our empirical dataset is constructed from a large-scale, in-house autonomous driving corpus gathered across geographically diverse regions in China. The raw recordings are organized hierarchically into continuous scenes, each comprising temporally aligned clips with synchronized multi-modal sensor streams and structured annotations. Every frame contains a timestamped front-view RGB image (704×1280 at 10 FPS) paired with synchronized vehicle state trajectories, sensor calibration profiles, high-precision LiDAR point clouds, per-frame semantic segmentation masks, and 3D agent bounding boxes parameterized by persistent tracking IDs, positions, and orientations in the ego-vehicle coordinate system. Together, these multi-modal streams provide the dense geometric, visual, and instance-level groundings necessary for multi-scale world model training.

D.2 Data Construction Pipeline

Raw driving logs inherently contain sensor anomalies, extreme motion, and misaligned annotations unsuitable for world-model training. To extract high-fidelity, structured training samples, we execute a systematic curation and annotation pipeline (Fig. 10).

We first enforce data quality through a three-stage filtering protocol. Stage I: Integrity Filtering discards sequences exhibiting dropped frames, exposure failures, or temporal discontinuities, preventing the generator from learning low-level sensor artifacts as scene dynamics. Stage II: Kinematic and Density Filtering mitigates motion blur and severe occlusion ambiguity by removing clips with average ego-velocities exceeding $45\,\mathrm{km/h}$ or heavily crowded scenes containing more than 20 annotated bounding boxes per frame. Stage III: Cross-Modal Alignment ensures precise geometric fidelity between spatial annotations and visual observations. We project HD-map elements and

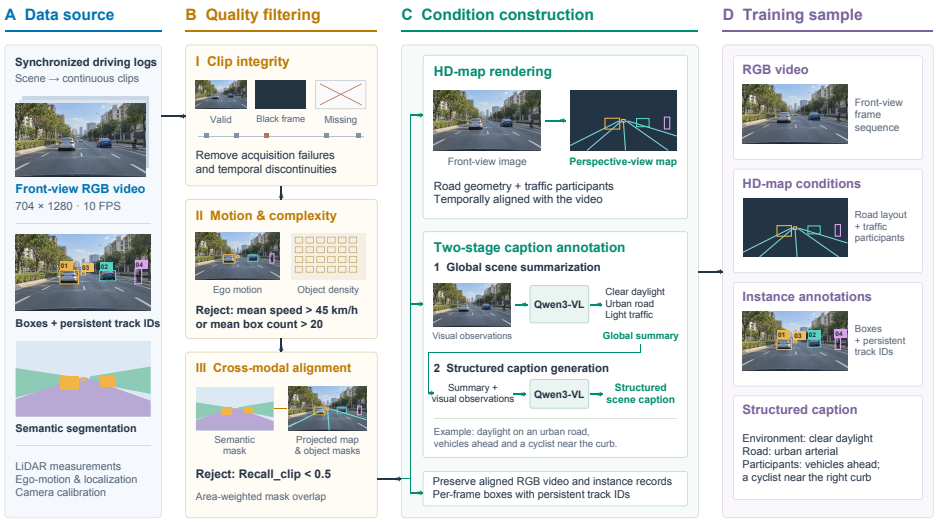

Figure 10: Dataset construction pipeline.

3D bounding boxes into the 2D image plane to compute their area-weighted overlap ($\mathrm{Recall}_{\mathrm{clip}}$) against corresponding semantic segmentation masks. Sequences with $\mathrm{Recall}_{\mathrm{clip}} < 0.5$ are rejected, systematically eliminating samples afflicted by projection drift or annotation lag.

For the curated sequences, we synthesize aligned conditioning signals to construct the final dataset. Structural Rendering and Captioning: We rasterize temporally synchronized HD maps to encode road topology and agent layouts. Concurrently, we generate dense textual descriptions using a two-stage prompting strategy with Qwen3-VL. The vision-language model first extracts global environmental priors (e.g., weather, illumination, traffic density) from the video. This summary subsequently conditions the generation of a comprehensive scene caption detailing fine-grained agent behaviors and ego-state dynamics. An example of the final generated description used as a video-generation prompt is shown in Figure 11.

Example of a Generated Video Prompt

Ego vehicle travels on a multi-lane urban roadway at night under clear conditions with artificial streetlight illumination; wet asphalt surface reflects light, lane markings are visible. Ahead in the center-left lane, a vehicle 10–15m distant drives forward with brake lights illuminated, indicating deceleration. A second vehicle in the center-right lane, 20m ahead, maintains forward motion. On the right sidewalk, 30m ahead, a pedestrian walks parallel to the road. A green traffic light at the upcoming intersection permits progression. A blue directional sign 40m ahead indicates routes to Yan'an Road and Changshu Road; a red vertical sign 15m ahead confirms Yan'an Road.

Figure 11: Example of a final textual description generated by the two-stage Qwen3-VL annotation pipeline and used as a video-generation prompt.

Ultimately, each training sample is standardized into four perfectly synchronized modalities: (i) a continuous front-view RGB video; (ii) temporally aligned HD-map sequences; (iii) per-frame 3D instance annotations parameterized by bounding-box coordinates and persistent track identifiers; and (iv) a hierarchical scene caption describing environmental contexts and salient traffic interactions.

D.3 Training and Evaluation Data

The curated training corpus comprises $121,794$ multi-modal driving clips ($\sim$1,000 cumulative hours) sampled at 10 FPS (704×1280 resolution). As illustrated in Fig. 12, clip durations follow a balanced temporal distribution. The corpus spans diverse geographic regions, urban topology, highway segments, and rural environments across China, encompassing varying illumination (day, dusk,

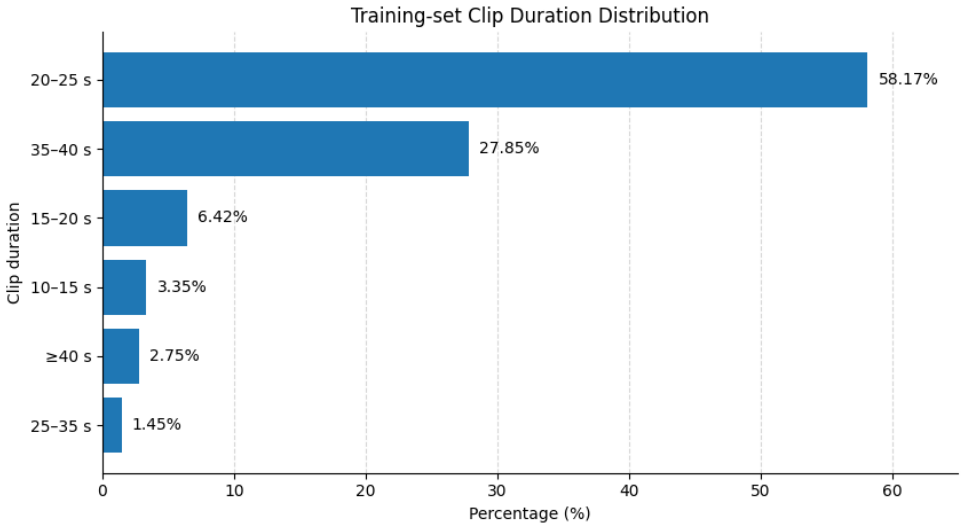

Figure 12: Distribution of clip durations in the training set.

night) and adverse weather conditions (clear, rain, fog). This environmental heterogeneity ensures robust exposure to complex road geometry, lighting transitions, and traffic density variations.

To rigorously assess long-horizon generation and consistency, we establish a standardized evaluation set of 350 curated clips. The evaluation protocol specifically targets challenging edge cases, including severe weather conditions, complex directional maneuvers (e.g., tight turns and unscripted lane changes), and dense interactions with vulnerable road users (pedestrians, cyclists). Furthermore, the evaluation suite isolates identity-sensitive sequences characterized by prolonged tracking occlusions and subsequent re-emergence, providing a stringent benchmark for measuring long-term instance identity persistence and scene stability.

Table 13: Summary of our Self-Collected Dataset.

Statistic	**Value**
Number of training clips	121,794
Total driving duration	$\sim$1,000 hours
Frame resolution	704×1280
Frame rate	10 FPS
Camera view	Front view
Geographic coverage	Diverse regions and cities in China
Road types	Urban / highway / rural
Time of day	Day / dusk / night
Weather conditions	Clear / rainy / foggy
Structured conditions	HD map + instance annotations + caption
Evaluation set	350 curated clips

D.4 Additional Dataset Details

Our self-collected corpus comprises driving sequences gathered across diverse urban, highway, and rural environments under varied diurnal and meteorological conditions. To preserve personal privacy, all identifiable elements—including human faces and license plates—undergo automated detection and rigorous spatial blurring prior to processing. Furthermore, location- and identity-associated metadata are anonymized in compliance with standard safety protocols.

The geolocated appearance database, $\mathcal{D}_{\text{geo}}$, indexes reference images using GPS coordinates structured via a spatial k-d tree. Image feature embeddings are pre-computed and cached to eliminate

redundant online extraction overhead, restricting spatial query latencies to under $5\,\mathrm{ms}$ during runtime retrieval.

Prior to training, raw logs undergo rigorous multi-stage filtering. Sequences containing sensor dropouts, black frames, or severe temporal discontinuities are systematically discarded. To maintain visual fidelity, we filter out clips featuring excessively high ego-velocities prone to motion blur, alongside over-crowded scenes exceeding dense participant thresholds. Finally, raw sensor streams—including LiDAR point clouds and camera parameters—are projected and rendered into synchronized, model-compatible conditioning representations.

E Evaluation and Metric Formulations

E.1 Quality-Horizon Evaluation

We evaluate long-horizon generation stability at three rollout durations, $\mathcal{H} = \{10, 30, 60\}\,\mathrm{s}$. Starting from the same initial conditioning, each method is rolled out using the same autoregressive chunk-wise generation protocol. For each target horizon $h \in \mathcal{H}$, we retain the generated sequence from the beginning of the rollout to h seconds and compute all metrics over the corresponding temporal range. The same evaluation samples, temporal horizons, backbone, sampler, and chunk configuration are used for all compared variants.

The ablation isolates the effect of world anchored memory. The baseline retains the structural HD-map conditioning of the backbone but removes the additional scene- and instance-memory readout branches, while our full model enables both. Thus, differences between the two variants reflect the contribution of persistent scene- and instance-level memory under an otherwise identical generation protocol.

We report three complementary measures. *Subject Coherence* evaluates the temporal appearance consistency of foreground traffic participants, while *Scene Similarity* measures the preservation of environment-level visual appearance over the rollout. *Fréchet Video Distance* (FVD) measures overall video distributional quality. Higher values are better for Subject Coherence and Scene Similarity, whereas lower values are better for FVD.

For a metric Q and method m, let $Q_m(h)$ denote its value evaluated over the rollout up to horizon h. We report

$$\mathcal{Q}_m = \{Q_m(h) \mid h \in \mathcal{H}\}\,, \tag{4}$$

which characterizes the evolution of generation quality as the autoregressive horizon increases. All methods are evaluated independently at the same three horizons using an identical protocol.

E.2 Scene Preservation Metrics

Scene Similarity (SS). Scene Similarity measures the visual consistency between the generated video and the ground-truth video within the scene regions. Let $\hat{\mathbf{I}}_t$ and $\mathbf{I}_t^{\mathrm{GT}}$ denote the generated and ground-truth frames at time t, respectively, and let $\mathbf{M}_t \in \{0,1\}^{H\times W}$ denote the corresponding scene mask. A frozen DINO image encoder ϕ extracts patch-level features, and the scene mask is average-pooled to the DINO patch grid to obtain patch weights $m_{t,p} \in [0,1]$.

The Scene Similarity of the entire video is defined as

$$\mathrm{SS} = \frac{\sum_{t=1}^{T}\sum_{p=1}^{P} m_{t,p}\,\cos\left(\phi_p(\hat{\mathbf{I}}_t), \phi_p(\mathbf{I}_t^{\mathrm{GT}})\right)}{\sum_{t=1}^{T}\sum_{p=1}^{P} m_{t,p}}, \tag{5}$$

where T is the number of video frames, P is the number of DINO patch tokens, and $\phi_p(\cdot)$ denotes the DINO feature of patch p. A higher SS indicates stronger scene-level appearance and structural consistency with the ground-truth video.

E.3 Identity Preservation Metrics

We evaluate instance identity preservation using appearance embeddings extracted from generated-video crops. For each visible instance i at frame t, we square-expand its ground-truth (GT) bounding

box by 10%, resize the resulting crop to 224×224, and extract an appearance embedding using a frozen DINOv2 ViT-B/14 model. The embedding is ℓ_2-normalized and denoted as $e_{t,i}$. All reported similarities are cosine similarities, and higher values indicate better identity preservation. We use a temporal stride of $S = 92$ frames, corresponding to the autoregressive chunk interval.

A track is considered valid for SC@kS if it contains at least one pair of frames separated by kS for which the instance remains continuously visible throughout the entire interval. For occlusion-based metrics, an occlusion event is defined as a maximal contiguous run of invisible frames that is strictly bounded by visible segments of the same track. The pre- and post-occlusion appearance prototypes are constructed only from the contiguous visible segments immediately adjacent to the corresponding occlusion event.

Long-range coherence (SC@kS). For each track, we compare appearance embeddings separated by a temporal gap of

$$\Delta = kS, \qquad k \in \{1, 2, 3\},$$

while requiring the instance to remain continuously visible throughout the interval.

Let $\mathcal{V}_{i,k}$ denote the set of valid starting frames for track i:

$$\mathcal{V}_{i,k} = \{t \,|\, \text{instance } i \text{ remains visible from } t \text{ through } t + kS\},$$

and let

$$\mathcal{T}_k = \{i \,|\, |\mathcal{V}_{i,k}| > 0\}$$

denote the set of valid tracks. We compute

$$\text{SC@}kS = \frac{1}{|\mathcal{T}_k|} \sum_{i \in \mathcal{T}_k} \frac{1}{|\mathcal{V}_{i,k}|} \sum_{t \in \mathcal{V}_{i,k}} e_{t,i}^\top e_{t+kS,i}, \qquad k \in \{1, 2, 3\}. \tag{6}$$

This metric measures long-range appearance coherence while excluding appearance changes associated with disappearance and re-emergence.

Occlusion and re-emergence. An occlusion event u is a maximal contiguous interval during which an instance is invisible and is preceded and followed by visible segments of the same track. Let $i(u)$ denote the corresponding instance.

For each event, $\mathcal{P}_u$ contains up to the last S frames from the contiguous visible segment immediately preceding the occlusion, and $\mathcal{Q}_u$ contains up to the first S frames from the contiguous visible segment immediately following re-emergence. We construct appearance prototypes on the two sides of the occlusion as

$$p_u^- = \text{Norm}\left(\frac{1}{|\mathcal{P}_u|} \sum_{t \in \mathcal{P}_u} e_{t,i(u)}\right), \qquad q_u^+ = \text{Norm}\left(\frac{1}{|\mathcal{Q}_u|} \sum_{t \in \mathcal{Q}_u} e_{t,i(u)}\right), \tag{7}$$

where $\text{Norm}(\cdot)$ denotes ℓ_2 normalization.

The re-emergence similarity for event u is then defined as

$$s_u = (p_u^-)^\top q_u^+. \tag{8}$$

Re-emergence similarity. We report a video-level macro-average to prevent videos containing many occlusion events from dominating the metric. Let $\mathcal{D}$ denote the set of videos containing at least one valid occlusion event, and let $\mathcal{U}_v$ denote the set of valid occlusion events in video v. We compute

$$\text{Re-emerg. Sim.} = \frac{1}{|\mathcal{D}|} \sum_{v \in \mathcal{D}} \frac{1}{|\mathcal{U}_v|} \sum_{u \in \mathcal{U}_v} s_u. \tag{9}$$

Re-emergence pass rate. We additionally report the fraction of valid occlusion events whose re-emergence similarity exceeds a GT-calibrated threshold τ. Let

$$\mathcal{U} = \bigcup_{v \in \mathcal{D}} \mathcal{U}_v$$

denote the set of all valid occlusion events. The pass rate is

$$\text{PassRate} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbf{1}[s_u \geq \tau]. \tag{10}$$

We set $\tau = 0.803$, corresponding to the median GT re-emergence similarity over all valid occlusion events in the evaluation set. The threshold is determined once from the GT videos and is fixed for all evaluated methods.

Occ-ID@mS. To analyze identity preservation under different occlusion durations, we group events according to their occlusion length L_u. For the m-th duration interval, we define

$$\mathcal{B}_m = \{u \,|\, (m-1)S < L_u \leq mS\}, \tag{11}$$

and compute

$$\text{Occ-ID@}mS = \frac{1}{|\mathcal{B}_m|} \sum_{u \in \mathcal{B}_m} s_u. \tag{12}$$

In our test set, we report Occ-ID@1S and Occ-ID@2S, corresponding to occlusion lengths of $(0, 92]$ and $(92, 184]$ frames, respectively. The Occ-ID@2S interval contains only three valid events, and no valid event has an occlusion duration longer than 184 frames.

Re-emergence similarity is macro-averaged at the video level, whereas Pass Rate and Occ-ID are averaged over individual occlusion events.

Average score. The reported Avg. score is the arithmetic mean of the seven metrics in Table 5: SC@1S, SC@2S, SC@3S, Re-emergence Similarity, Re-emergence Pass Rate, Occ-ID@1S, and Occ-ID@2S.

The GT videos are used only to determine track identities, visibility, occlusion events, bounding boxes, and the calibration threshold τ. All appearance similarities reported for each evaluated method are computed exclusively from the generated videos.

E.4 RL Evaluation Metrics

Subject Fidelity. We measure whether the conditioned dynamic subjects are detectable in the generated video with a GT-gated detector protocol. Conditioning boxes (per-frame tight AABBs of the conditioned pedestrians, cyclists, cars and vehicles) are taken from the ground-truth annotations. A YOLO26s detector trained on BDD100k is run on every frame of the generated video; each conditioning box is greedily matched to the highest-IoU same-class detection. To exclude annotation and visibility ambiguity from the denominator, the identical detection-and-matching procedure is run on the ground-truth video, and only instances the detector recovers in the GT frames count as valid.

Physical Plausibility. We use Qwen3.5-9B as a no-reference VLM judge to evaluate pedestrian motion plausibility. Each video is evaluated with two binary questions covering motion anatomy and limb integrity. We use the normalized probability of "yes" at the first generated token rather than a sampled binary response. Specifically, the probability is computed by renormalizing the logits over the predefined yes/no token variants. The final score is the mean $P(\text{yes})$ across the two questions. Only videos with clearly visible pedestrians are included in this evaluation.

Physical-Plausibility Prompts

Motion Anatomy: Do all walking or running pedestrians show perfectly natural, anatomically plausible limb movement throughout the entire video, with zero unnatural joint motion and zero abnormal limb trajectories? *(yes or no)*

Limb Integrity: For moving pedestrians, do their arms and legs remain clearly distinguishable throughout the motion, without limbs merging into the body, merging with each other, disappearing, or becoming visually entangled? *(yes or no)*

F Limitations and Future Work

Despite its strong long-horizon consistency, WAM-Sim exhibits three primary limitations. First, memory retrieval latency and attention memory footprint scale linearly with the number of tracked

foreground instances, introducing computational overhead in hyper-dense traffic scenes containing over 100 concurrent agents. Second, the write-once instance appearance bank assumes rigid visual properties; consequently, extreme non-rigid state transitions (e.g., severe vehicle damage during collisions or drastic open/close deformations of commercial truck beds) are constrained by the initial-observation tokens. Third, background scene retrieval depends on geolocated reference availability, which may degrade to ungrounded autoregressive generation in unmapped or highly dynamic construction zones.

To address these challenges, future iterations will pursue three key research directions: 1) Integrating explicit 3D visual representations, such as online-updated Gaussian Splatting or neural radiance fields, into the memory substrate to enable unified multi-view spatial consistency and continuous geometry-aware retrieval under counterfactual maneuvers. 2) Formulating a parametric write-update memory mechanism that selectively incorporates visual state changes across significant illumination or topological shifts while preserving fundamental instance identity invariants. 3) Developing dynamic instance-grouping and memory-compression protocols to bound memory complexity in extreme ultra-dense urban traffic scenarios.

G Ethics, Broader Impact, and Reproducibility Assurance

WAM-Sim advances autonomous driving safety by enabling controllable, zero-cost edge-case simulation over extended horizons. While high-fidelity video generation presents potential dual-use risks regarding synthetic media misuse, these concerns can be effectively mitigated through digital watermarking protocols and restricted access frameworks for generative models.

To ensure full reproducibility, we release our complete training and evaluation code, model checkpoints, and dataset processing pipelines. Detailed architectural configurations, optimization hyperparameters, and hardware environments are thoroughly documented in Appendices C and D.