SIMWAM: A SIMPLE WORLD ACTION MODEL FOR END-TO-END AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

In autonomous driving, World-Action Models (WAMs) have improved end-to-end planning by transferring video dynamics priors to action prediction, but many still couple planning with future-video generation at inference, incurring substantial computational overhead. We present **SimWAM**, a simple yet effective WAM that leverages future-video prediction solely as a training-time supervision signal. It co-trains a pretrained video expert and a lightweight action expert with joint flow matching. An isolated attention mask keeps action prediction independent of future frames, allowing trajectory prediction without future-frame generation at inference. This design supports multiple pretrained video backbones and independent action-expert scaling within a shared attention interface, while preserving the joint learning objective. Moreover, we apply reinforcement learning to optimize a compositional driving reward beyond trajectory imitation. Experiments show that SimWAM achieves 91.9 PDMS on NAVSIM with a favorable trade-off between accuracy and latency among world-model-based planners, while transferring zero-shot to nuScenes. It also achieves competitive planning accuracy on WOD-E2E and PhysicalAI-Autonomous-Vehicles. These results position SimWAM as a plain yet solid baseline for efficient autonomous driving.

1 INTRODUCTION

End-to-end autonomous driving (AD) (Codevilla et al., 2018) maps raw sensor observations directly to a planned trajectory with a unified network. Joint optimization removes hand-crafted interfaces and reduces error propagation in the classical perception, prediction, and planning pipeline (Pomerleau, 1988; Bojarski et al., 2016). Although recent end-to-end planners (Hu et al., 2023; Jiang et al., 2023; Liao et al., 2025) have steadily improved planning accuracy, they remain largely black-box mappings from observations to trajectories. In such models, traffic semantics, user intent, and scene dynamics remain only implicitly encoded within the learned mapping.

Vision-Language-Action (VLA) models (Kim et al., 2024; Black et al., 2025; Li et al., 2026d;a; Fu et al., 2025) address the semantic limitation by adapting pretrained vision-language models to driving. Their world knowledge and reasoning improve scene understanding and connect trajectory generation with user intent. Many driving VLAs (Zhao et al., 2025a; Zhou et al., 2026; Wang et al., 2025c) further produce an explicit rationale before trajectory prediction, improving interpretability in complex and instruction-conditioned scenarios. Recent methods (Zeng et al., 2025; Tan et al., 2026; Peng et al., 2026) introduce future-scene generation or latent reasoning to strengthen spatiotemporal understanding. However, these components remain loosely coupled with action prediction and often require additional training stages or sequential inference. Motion and temporal evolution are still modeled only indirectly, calling for an explicit prior over world dynamics.

World models meet this demand by furnishing an explicit prior over how the environment evolves under motion. Building on this principle, recent World-Action Models (WAMs) in embodied intelligence, such as DreamZero (Ye et al., 2026) and LingBot-VA (Li et al., 2026b), jointly predict future observations and actions through pretrained video-generation backbones. This world-action paradigm has recently been adopted in autonomous driving. DriveLaW (Xia et al., 2026) and DriveWAM (Shi et al., 2026) co-train a video predictor and a planner, allowing anticipated scene dynamics to inform trajectory generation. Nevertheless, existing driving WAMs commonly follow an *imagine-then-act* pipeline in which the planner conditions its output on explicit future visual

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.

prediction. This reliance on future synthesis incurs substantial inference overhead, as reflected in Fig. 1.

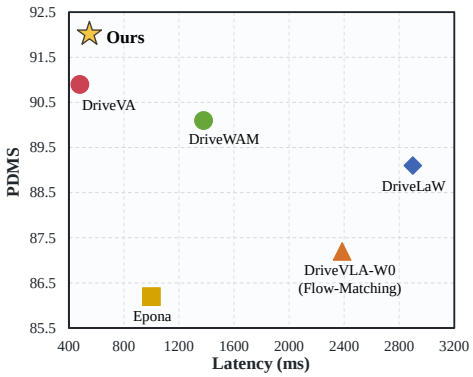


Figure 1: SimWAM achieves the best PDMS among world-model-based planners, with a favorable trade-off between accuracy and latency on one A100 GPU.

Indeed, Fast-WAM (Yuan et al., 2026), a world-action model from embodied intelligence, shows that explicit future synthesis is unnecessary for effective world-action learning, as video co-training benefits action prediction primarily through *training-time* representation learning rather than *test-time* imagination. Building on this insight, we introduce **SimWAM**, a plain yet effective World-Action Model that uses video generation as a training signal while retaining direct trajectory prediction at inference. SimWAM jointly trains a pretrained video expert and a lightweight action expert with flow matching. A simple isolated attention mask prevents the action expert from accessing future frames, so inference bypasses explicit future-frame prediction while the action expert still leverages the learned traffic-dynamics prior. Building on this mask, more advanced video generators could further improve the learned prior without touching the action expert or the inference pipeline. Furthermore, we reformulate the deterministic flow ODE as a stochastic SDE and reinforce the action expert with GRPO (Guo et al., 2025; Liu et al., 2025), enabling diverse maneuver exploration and direct optimization of a compositional driving reward. Rather than claiming algorithmic superiority, this work establishes a simple and solid WAM baseline for exploring the potential of generic video models in AD.

Experiments on three real-world benchmarks validate this simple design. On NAVSIM (Dauner et al., 2024), SimWAM achieves 91.9 PDMS and offers a favorable trade-off between planning performance and inference latency among current world-model-based planners, as shown in Fig. 1. The same NAVSIM-trained model further transfers zero-shot to nuScenes (Caesar et al., 2020) without fine-tuning, demonstrating cross-domain generalization. On WOD-E2E and PhysicalAI-Autonomous-Vehicles, SimWAM similarly delivers competitive planning accuracy. We hope SimWAM will serve as a strong and practical baseline for efficient world-action modeling in AD.

The advantages of SimWAM arise from three aspects. First, SimWAM transfers the traffic-dynamics prior of a pretrained video expert to the action expert via joint co-training. Second, thanks to the isolated attention mask, the action expert remains independent of future-frame representations, allowing efficient inference without explicitly generating future frames. Third, the decoupled architecture seamlessly accommodates more advanced video generators without modifying the action expert.

2 Related Work

2.1 Vision-Language-Action Models for Autonomous Driving

End-to-end autonomous driving integrates perception, prediction, and planning in a unified framework, as demonstrated by UniAD (Hu et al., 2023) and VAD (Jiang et al., 2023), but remains largely dependent on expert trajectory supervision with limited semantic reasoning. Vision-Language-Action (VLA) models (Li et al., 2026d; Fu et al., 2025; Zhou et al., 2025; Liu et al., 2026c) leverage pretrained vision-language representations for semantic reasoning, with AutoVLA (Zhou et al., 2026) and ORION (Fu et al., 2025) further combining reasoning with trajectory generation. FutureSightDrive (Zeng et al., 2025) and ExploreVLA (Sheng et al., 2026) introduce future image generation for scene evolution modeling. In contrast, SimWAM transfers motion priors from a pretrained video generator to a lightweight action expert for direct trajectory prediction.

2.2 World-Action Models for Autonomous Driving

World-Action Models (Assran et al., 2025; Wang et al., 2025b; Agarwal et al., 2025) jointly learn action prediction and image generation to capture object motion, physical interactions and future scene evolution, as demonstrated in robotics by DreamZero (Ye et al., 2026) and LingBot-VA (Li et al., 2026b). In autonomous driving, earlier world models such as DriveDreamer (Wang et al.,

2024) and HERMES (Zhou et al., 2025) mainly predicted or generated future driving scenes, while more recent studies (Li et al., 2026c; Shi et al., 2026) integrate visual world modeling with trajectory planning. Epona (Zhang et al., 2025c) jointly predicts future videos and trajectories via autoregressive diffusion, whereas DriveLaW (Xia et al., 2026) conditions a diffusion planner on latents produced by a video generator. They couple planning with future visual generation at inference. SimWAM instead uses future-video generation only as training supervision and predicts trajectories directly.

2.3 Reinforcement Learning for Autonomous Driving

Imitation learning trains autonomous driving policies to reproduce expert trajectories, but this objective confines learning to demonstrated behavior and indirectly reflects overall driving quality. Reinforcement learning provides a complementary refinement that directly optimizes driving policies with task-level rewards. CarPlanner (Zhang et al., 2025b) uses expert-guided rewards to improve large-scale trajectory planning. Raw2Drive (Yang et al., 2025) refines driving policies with raw sensor inputs and privileged world models. Recent studies (Zhou et al., 2026; Li et al., 2026d; Fu et al., 2026; Sheng et al., 2026) have further introduced reinforcement learning into Vision-Language-Action driving models. MindDrive (Fu et al., 2026) applies online reinforcement learning to refine high-level decisions using feedback from their induced driving trajectories. CritiqueDriveVLM (Liu et al., 2026b) refines driving reasoning with a verifier and distills this capability into an efficient reasoning model. SimWAM instead applies reward optimization directly to its continuous trajectory-generating action expert.

3 Preliminary

Flow matching. Rectified flow (Lipman et al., 2023; Liu et al., 2023) interpolates a clean target x with Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ as $x_\tau = (1-\tau)\, x + \tau\, \epsilon$ ($\tau \in [0, 1]$). The network v_θ learns the target velocity $\epsilon - x$ under conditioning c:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{x,\epsilon,\tau}\big[\, \|v_\theta(x_\tau, \tau, c) - (\epsilon - x)\|_2^2 \,\big]. \tag{1}$$

Sampling integrates the probability-flow ODE $\mathrm{d}x_\tau = v_\theta(x_\tau, \tau, c)\, \mathrm{d}\tau$ from noise (τ=1) to data (τ=0), so that each noise sample is transported along a deterministic trajectory to one generated sample.

From ODE to SDE. The deterministic ODE generates a single sample per noise input and lacks a tractable transition density, which restricts exploration and precludes policy-gradient optimization. To introduce stochasticity into this deterministic framework, SDE-based GRPO methods (Liu et al., 2025) adopt sampling with the same marginal distributions as the original ODE:

$$\mathrm{d}x_\tau = \Big[v_\theta(x_\tau, \tau) + \tfrac{\sigma_\tau^2}{2\tau}\big(x_\tau + (1-\tau)\, v_\theta(x_\tau, \tau)\big)\Big]\mathrm{d}\tau + \sigma_\tau\, \mathrm{d}w, \qquad \sigma_\tau = \eta\sqrt{\tfrac{\tau}{1-\tau}}, \tag{2}$$

where $\mathrm{d}w$ is a Wiener increment and η controls the noise scale. Each Euler-Maruyama step of size $\Delta\tau$ yields an isotropic Gaussian transition $\pi_\theta(x_{\tau-\Delta\tau} \mid x_\tau) = \mathcal{N}\big(\mu_\theta(x_\tau, \tau),\, \sigma_\tau^2 \Delta\tau\, I\big)$ with tractable log-likelihoods for importance sampling, in which $\mu_\theta(x_\tau, \tau)$ is the drift-induced mean of the update and reduces to the deterministic ODE step as η vanishes.

4 Method

We present SimWAM as a strong and practical baseline for end-to-end autonomous driving rather than a pursuit of architectural innovation. As illustrated in Fig. 2, the design stays deliberately plain, with a pretrained video expert and a lightweight action expert co-trained by joint flow matching under an isolated attention mask that keeps trajectory prediction free of future frames at inference.

4.1 Model Architecture

Problem formulation. We consider end-to-end trajectory planning from a front-camera observation o_t, the ego state s_t containing velocity, acceleration, and yaw rate, and a navigation command l. The planner predicts an ego trajectory $a_{t+1:t+H} = (a_{t+1}, \ldots, a_{t+H})$ in the ego-vehicle coordinate frame, where H is the number of waypoints and each waypoint $a_i = (x_i, y_i, \psi_i)$ specifies the planned

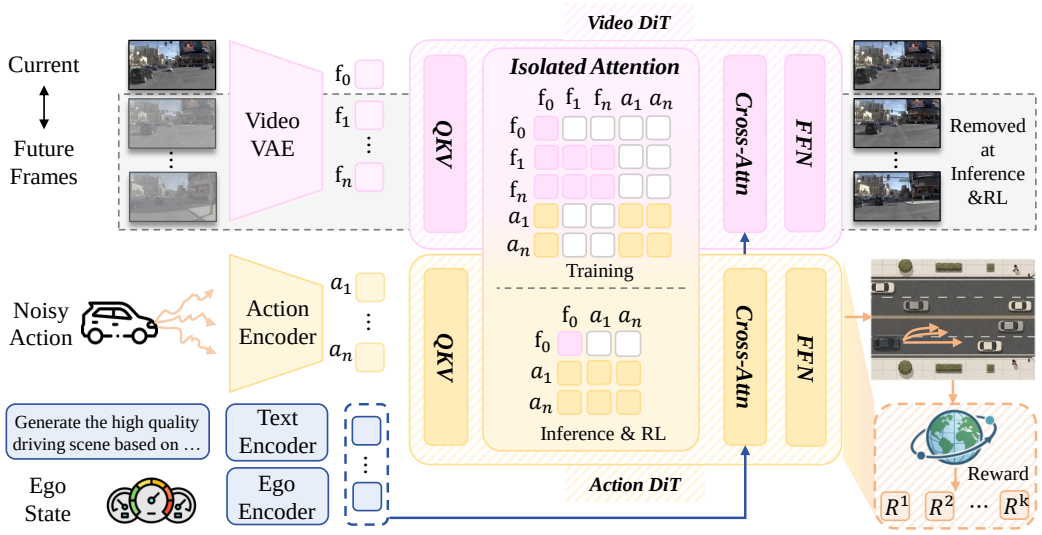


Figure 2: **Overview of SimWAM.** During training, the video and action DiTs are jointly optimized for future-frame generation and trajectory prediction via shared attention, while the isolated mask prevents the action tokens from accessing future-frame tokens. During inference and reinforcement learning, the model directly predicts trajectories without explicitly generating future frames.

position and heading. Existing driving WAMs (Xia et al., 2026; Shi et al., 2026) commonly follow an imagine-then-act paradigm, factorizing the policy as:

$$p_\theta(a_{t+1:t+H},\, z_{t+1:t+N} \mid o_t, s_t, l) = p_\theta(z_{t+1:t+N} \mid o_t, s_t, l)\, p_\theta(a_{t+1:t+H} \mid o_t, s_t, l, z_{t+1:t+N}), \quad (3)$$

which conditions trajectory generation on the predicted latents $z_{t+1:t+N}$ of N future frames, adding video-generation overhead to planning. Our SimWAM departs from this paradigm by absorbing the traffic-dynamics prior from future-video supervision during co-training, while retaining a plain and direct policy interface without future-scene generation at inference, given by:

$$p_\theta(a_{t+1:t+H} \mid o_t, s_t, l) = p_\theta\big(a_{t+1:t+H} \mid z(o_t), s_t, l\big), \quad (4)$$

where $z(o_t)$ is the representation produced from the current observation.

Video expert. We initialize the video Diffusion Transformer (DiT) (Peebles & Xie, 2023) from Wan2.2-5B (Wan et al., 2025) and use its accompanying video VAE (Kingma & Welling, 2014) and T5 text encoder (Raffel et al., 2020). The current frame serves as the clean condition, and the navigation command l is first encoded by the T5 text encoder and then injected into the DiT via cross-attention, while the N future frames are noised and reconstructed with flow matching. This video-generation objective endows the action expert with a traffic-dynamics-aware representation by forcing the model to anticipate how traffic evolves.

Action expert. The action expert is a lightweight Diffusion Transformer with hidden size d_a=1024. Conditioned on $c = \{z(o_t), s_t, l\}$ with the ego state s_t embedded by a small MLP, it predicts the trajectory velocity field $v_\theta(a^\tau_{t+1:t+H}, \tau, c)$ via flow matching. Integrating this velocity field along the probability-flow ODE from Gaussian noise at τ=1 to τ=0 then yields the planned trajectory through the same flow-matching interface as the video expert.

Isolated attention mask. To exploit future-video supervision during training without paying for future-frame synthesis at inference, we adopt the isolated attention mask of Fig. 2. Specifically, the future-frame tokens $z_{t+1:t+N}$ and the action tokens both attend to the current observation $z(o_t)$ while remaining mutually invisible. Consequently, the action expert uses current-frame visual representations $z(o_t)$, together with the ego state and navigation command, without accessing future-frame tokens. Moreover, the same decoupling later allows reinforcement learning to optimize trajectories without any future-frame rollout (§4.2).

Co-training. The two experts interact only through shared attention and retain their original architectures. Joint flow matching over video and trajectory modalities allows future-scene prediction to shape the observation representation used for planning. The joint objective is defined as:

$$\mathcal{L} = \mathcal{L}^{\text{act}}_{\text{FM}} + \lambda\, \mathcal{L}^{\text{vid}}_{\text{FM}}, \quad (5)$$

where $\mathcal{L}_{\text{FM}}^{\text{act}}$ and $\mathcal{L}_{\text{FM}}^{\text{vid}}$ instantiate Eq. 1 on the action trajectory $a_{t+1:t+H}$ and the future-frame latents $z_{t+1:t+N}$, and λ balances the two terms.

4.2 Reinforcement Learning

Imitation learning fits logged trajectories without directly optimizing the specific driving reward. We therefore refine the action expert with reinforcement learning on a task-level driving reward. As in Sec. 3, we use the SDE in Eq. 2 for stochastic rollouts. Its Gaussian transitions provide exploration and tractable per-step conditional log-likelihoods for policy optimization. To keep the method simple, we adapt the SDE to the trajectory output space with minimal stochasticity. Empirically, SDE integration suits this space better in the later denoising steps than in the earlier ones. We therefore activate it only on the last three steps $\mathcal{W} = \{K-3, K-2, K-1\}$ of the K-step chain, whose updates are given by:

$$a^{(k+1)} = \mu_{\theta,k} + \sigma_k \, \xi_k, \quad \sigma_k = \eta \sqrt{\frac{\tau_k \left(\tau_k - \tau_{k+1}\right)}{1 - \tau_k}}, \quad \xi_k \sim \mathcal{N}(0, I), \quad k \in \mathcal{W}, \tag{6}$$

where $\mu_{\theta,k}$ is the SDE mean at step k under schedule τ_k. For each scenario, we score G candidates and obtain rewards r_i. We compute the group-relative advantage $\hat{A}_i = (r_i - \bar{r})/(\sigma_r + \zeta)$, where $\bar{r}$ and σ_r are the group mean and standard deviation and ζ is a small constant avoiding division by zero. The clipped surrogate objective (Shao et al., 2024; Guo et al., 2025) is as follows:

$$\begin{aligned} \mathcal{L}_{\text{RL}} &= -\mathbb{E}\left[\min\left(\frac{\pi_\theta(a_i^{(k+1)}|a_i^{(k)},c)}{\pi_{\text{old}}(a_i^{(k+1)}|a_i^{(k)},c)}\hat{A}_i,\ \text{clip}\left(\frac{\pi_\theta(a_i^{(k+1)}|a_i^{(k)},c)}{\pi_{\text{old}}(a_i^{(k+1)}|a_i^{(k)},c)}, 1-\delta, 1+\delta\right)\hat{A}_i\right)\right] + \beta\mathcal{L}_{\text{anchor}}, \\ \mathcal{L}_{\text{anchor}} &= \mathbb{E}_{i,k}\left[\frac{1}{H\sum_d w_d}\sum_{h=1}^{H}\sum_{d=1}^{3} w_d\left(\mu_{\theta,ikhd} - \mu_{\text{ref},ikhd}\right)^2\right], \end{aligned} \tag{7}$$

where the expectations run over candidates i and steps $k \in \mathcal{W}$, and δ is the clip range. The anchor regularizes the SDE mean toward the frozen imitation policy at identical states, with coordinate weights $w = (0.2, 1, 1)$ leaving the longitudinal coordinate freer to adapt.

4.3 Flexibility of SimWAM

The structural simplicity of SimWAM naturally yields flexibility in architecture and model scale. The two experts are parameterized independently and exchange information solely through the attention stream, so their architectures and capacities can be adjusted separately within this interface.

Video expert flexibility. Thanks to this simple interface, SimWAM can seamlessly accommodate different pretrained video experts. The action expert operates on current-frame representations supplied by the video expert rather than future-frame tokens or decoded predictions. Future-video supervision shapes these representations during co-training, while the video expert continues to encode the current observation at inference without generating future frames. Replacing the video backbone with a newer or more driving domain-relevant model thus requires no redesign of the action expert or trajectory objective. In this sense, SimWAM can readily benefit from advances in video generation while preserving the same video-action interface.

Scale flexibility. Within the shared attention interface, the action expert can be scaled independently of the video expert. The video expert supplies current-frame visual features shaped by video pretraining and co-training, while the width of the action expert controls the trade-off between planning capacity and computational cost. This design accommodates different computational budgets without changing the video expert or the training objective.

5 Experiments

5.1 Experimental Setup

Datasets and benchmarks. We evaluate SimWAM on three real-world driving benchmarks. NAVSIM (Dauner et al., 2024) provides non-reactive pseudo-simulation with the PDMS metric.

Table 1: Comparison with state-of-the-art planners on the NAVSIM `navtest` benchmark. C denotes camera and L denotes LiDAR.

Method	Reference	Sensors	NC↑	DAC↑	EP↑	TTC↑	C↑	PDMS↑
Human Agent	-	-	100.0	100.0	87.5	100.0	99.9	94.8
Traditional E2E planners								
UniAD (Hu et al., 2023)	CVPR'23	6×C	97.8	91.9	78.8	92.9	100.0	83.4
TransFuser (Chitta et al., 2022)	TPAMI'22	3×C+L	97.7	92.8	79.2	92.8	100.0	84.0
DiffusionDrive (Liao et al., 2025)	CVPR'25	3×C+L	98.2	96.2	82.2	94.7	100.0	88.1
WoTE (Li et al., 2025)	ICCV'25	3×C+L	98.5	96.8	81.9	94.9	99.9	88.3
SeerDrive (Zhang et al., 2025a)	NeurIPS'25	3×C+L	98.4	97.0	83.2	94.9	99.9	88.9
VLM-based planners								
AutoVLA (Zhou et al., 2026)	NeurIPS'25	3×C	98.4	95.6	81.9	98.0	99.9	89.1
ReCogDrive (Li et al., 2026d)	ICLR'26	1×C	97.9	97.3	**87.3**	94.9	100.0	90.8
DriveVLA-W0 (Flow-Matching) (Li et al., 2026c)	ICLR'26	1×C	98.4	95.3	80.9	95.2	100.0	87.2
DriveVLA-W0 (Query-Base) (Li et al., 2026c)	ICLR'26	1×C	98.7	**99.1**	83.3	95.3	99.3	90.2
ExploreVLA (Sheng et al., 2026)	ECCV'26	1×C	98.8	98.4	83.5	96.5	99.9	90.4
SGDrive (Li et al., 2026a)	CVPR'26	1×C	98.6	97.8	85.8	96.2	100.0	91.1
World-model-based planners								
Epona (Zhang et al., 2025c)	ICCV'25	1×C	97.9	95.1	80.4	93.8	99.9	86.2
PWM (Zhao et al., 2025b)	NeurIPS'25	1×C	98.6	95.9	81.8	95.4	100.0	88.1
DriveLaW (Xia et al., 2026)	CVPR'26	1×C	99.0	97.1	81.3	96.7	100.0	89.1
DriveVA (Liu et al., 2026a)	ECCV'26	1×C	**99.2**	97.5	83.5	**98.7**	100.0	90.9
DriveWAM (Shi et al., 2026)	arXiv'26	1×C	98.3	98.1	84.3	95.2	100.0	90.1
SimWAM (**ours**)	-	1×C	98.5	99.1	86.5	96.1	99.8	**91.9**

Table 2: Future-video generation quality on `navtest`. All methods use a 4 s horizon at 2 Hz.

	SVD	DrivingGPT	Uni-World VLA	PWM	ForgeDrive	DriveDreamer-Policy	SimWAM (**ours**)
FVD↓	227.5	142.6	141.8	86.0	69.2	53.6	**36.3**

Table 3: Open-loop planning on the WOD-E2E benchmark. ADE denotes the average displacement error, and RFS the Rater Feedback Score. $*$ indicates reinforcement learning.

(a) Validation split.

Method	ADE@3s↓	ADE@5s↓	RFS↑
Human Driver	-	-	8.13
VAD (Jiang et al., 2023)	3.19	5.81	4.45
UniAD (Hu et al., 2023)	6.50	10.81	5.78
RAP-DINO (Feng et al., 2026)	0.97	2.20	7.91
SimWAM (**ours**)	0.98	2.29	7.99
SimWAM (**ours**)*	0.98	2.27	**8.29**

(b) Test split.

Method	ADE@3s↓	ADE@5s↓	RFS↑
UniPlan (Liao et al., 2025)	1.31	2.99	7.78
dVLM-AD (Ma et al., 2026)	1.29	3.02	7.63
HMVLM (Wang et al., 2025a)	1.33	3.07	7.74
AutoVLA* (Zhou et al., 2026)	1.35	2.96	7.56
NoRD* (Rawal et al., 2026)	1.25	-	7.71
SimWAM (**ours**)	1.22	2.69	7.77
SimWAM (**ours**)*	1.21	2.68	7.84

The Waymo Open Dataset end-to-end benchmark (WOD-E2E) (Xu et al., 2026) focuses on long-tail scenarios and is scored by the Rater Feedback Score (RFS) and the average displacement error (ADE). PhysicalAI-Autonomous-Vehicles (Wang et al., 2025c) (PhysicalAI-AV) offers large-scale real-world driving logs evaluated with the ADE and Final Displacement Error (FDE).

Implementation details. The video expert is initialized from Wan2.2-5B (Wan et al., 2025) and the action expert is a lightweight 1024-wide DiT. For imitation learning (IL), we train for 100 epochs on NAVSIM, 30 epochs on WOD-E2E, and 15 epochs on PhysicalAI-AV. We use front-camera images at resolutions of 384×672 on NAVSIM and 512×480 on WOD-E2E. We subsequently apply reinforcement learning (RL) on NAVSIM and WOD-E2E, while PhysicalAI-AV uses imitation learning only. RL optimizes only LoRA adapters on the action expert, with the SDE restricted to the last three steps. See more details in Appendix B.

5.2 Main Results

NAVSIM. As shown in Tab. 1, our method achieves 91.9 PDMS on the test split with a single front camera, surpassing recent VLM-based and world-model-based end-to-end planners. It exceeds SGDrive (Li et al., 2026a) by 0.8 points and ExploreVLA (Sheng et al., 2026), a VLA with future image generation, by 1.5 points. Notably, our future-synthesis-free model further bests the imagine-then-act planners DriveLaW (Xia et al., 2026) and DriveVA (Liu et al., 2026a) by 2.8 and 1.0 points.

Table 4: Comparison on the 1,000-clip test subset of the PhysicalAI-Autonomous-Vehicles benchmark (Shi et al., 2026). Params. denotes the number of model parameters. $*$ indicates evaluation with the released checkpoint, which supports prediction up to 3 s only.

Method	Source	Params.	ADE@3s↓	FDE@3s↓	ADE@4s↓	FDE@4s↓
VaVAM$*$ (Bartoccioni et al., 2025)	Valeo	1.3B	2.31	4.32	-	-
Alpamayo-1.5 (Wang et al., 2025c)	NVIDIA	10B	0.80	2.31	1.44	4.18
DriveWAM (Shi et al., 2026)	-	5B + 8B	0.47	1.35	0.83	2.47
SimWAM (**ours**)	-	6B	0.40	1.08	0.69	1.96

Table 5: Zero-shot generalization on the nuScenes open-loop planning benchmark. $*$ represents only using the front camera as input.

Method	Finetune	Input	Auxiliary Supervision	L2 (m)↓				Collision Rate (%)↓			
				1 s	2 s	3 s	Avg.	1 s	2 s	3 s	Avg.
ST-P3 (Hu et al., 2022b)	✓	Camera	Map&Box&Depth	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71
UniAD (Hu et al., 2023)	✓	Camera	Map&Box&Motion	0.48	0.96	1.65	1.03	0.05	0.17	0.71	0.31
VAD-Base (Jiang et al., 2023)	✓	Camera	Map&Box&Motion	0.54	1.15	1.98	1.22	0.04	0.39	1.17	0.53
GenAD (Zheng et al., 2024)	✓	Camera	Map&Box&Motion	0.36	0.83	1.55	0.91	0.06	0.23	1.00	0.43
Epona (Zhang et al., 2025c)	✓	Camera*	None	0.61	1.17	1.98	1.25	0.01	0.22	0.85	0.36
DriveVA (Liu et al., 2026a)	✗	Camera*	None	0.33	**0.76**	**1.43**	**0.84**	0.00	0.07	0.12	0.06
DriveWAM (Shi et al., 2026)	✗	Camera*	None	**0.28**	0.81	1.80	0.96	0.00	0.05	0.14	0.06
SimWAM (**ours**)	✗	Camera*	None	0.29	0.82	1.77	0.96	0.00	**0.03**	**0.11**	**0.05**

Moreover, the co-trained video expert achieves the lowest FVD of 36.3 in Tab. 2, compared with 53.6 for the next-best method. These results show that SimWAM combines strong planning performance with high-quality video generation, while keeping trajectory inference independent of future synthesis. More results appear in Appendix C.

WOD-E2E. Tab. 3 compares open-loop planning on the WOD-E2E validation and test splits. Although reinforcement learning is performed on the validation split, the imitation-trained model alone already surpasses the strongest baseline RAP-DINO (Feng et al., 2026) there with an RFS of 7.99. On the held-out Waymo test split, the imitation-trained model already surpasses all compared methods in ADE, and reinforcement learning generalizes to this split as well, improving RFS from 7.77 to 7.84. Notably, SimWAM achieves this without language reasoning or test-time future synthesis, highlighting the strength of combining joint video co-training with reinforcement learning.

PhysicalAI-AV. We further examine open-loop trajectory prediction on PhysicalAI-Autonomous-Vehicles in Tab. 4, which reports only the imitation-trained model. Trained on only 65K samples, SimWAM attains the lowest ADE and FDE at both the 3s and 4s horizons, surpassing VaVAM (Bartoccioni et al., 2025) and Alpamayo-1.5 (Wang et al., 2025c) despite their far larger training corpora. More notably, DriveWAM (Shi et al., 2026) builds on the same video expert but relies on test-time future-frame synthesis. SimWAM outperforms it without any explicit future-video generation.

Zero-shot generalization on nuScenes. We further evaluate the NAVSIM-trained SimWAM on the nuScenes (Caesar et al., 2020) open-loop benchmark without fine-tuning. As shown in Tab. 5, despite using neither nuScenes supervision nor auxiliary annotations, SimWAM achieves an average L2 error and collision rate comparable to or better than those of strong zero-shot baselines, indicating that its learned dynamics-aware representation transfers beyond the training benchmark.

Overall, despite being deliberately plain and forgoing bells and whistles such as language reasoning or test-time future synthesis, SimWAM consistently matches or even surpasses substantially more complex planners across diverse benchmarks, clearly demonstrating that a simple world-action design is already remarkably robust and generalizable for end-to-end driving.

5.3 Ablation Studies And Analysis

Unless otherwise noted, all ablations report the imitation-trained world-action model on the NAVSIM `navtest` split, and all inference latencies are measured on a single NVIDIA A100 GPU.

Analysis on the attention mask. The attention pattern determines how information flows between the two experts, and we compare three alternatives in Tab. 6. Both bidirectional and action→video attention tie action prediction to future video tokens, forcing the model to instantiate future-frame

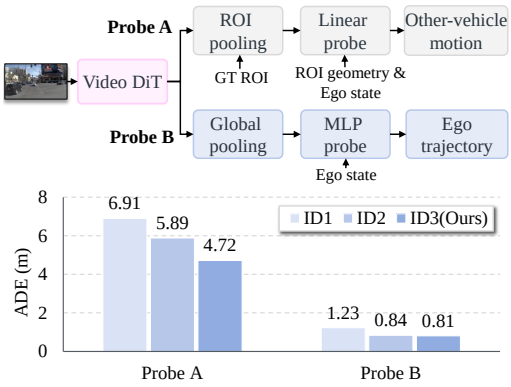


Figure 3: Probe-based representation analysis on the current-frame features of the video expert.

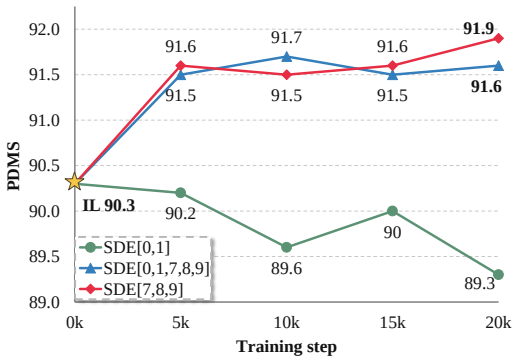


Figure 4: Ablation on the placement of the SDE steps during RL.

Table 6: Ablation on attention mask.

Mask	NC	DAC	EP	TTC	PDMS	Latency (ms)↓
Bidirectional	98.4	**98.0**	**84.7**	95.1	90.2	1512
Action→video	98.5	97.8	84.3	95.5	90.1	1499
Isolated	**98.6**	98.0	84.0	**95.8**	**90.3**	**518**

Table 7: Video expert flexibility. RL in parentheses.

Video model	NC	DAC	EP	TTC	PDMS
LTX-Video	98.2 (98.6)	97.2 (98.6)	83.0 (85.5)	94.4 (95.3)	88.7 (90.9)
Wan2.1-1.3B	98.6 (98.6)	**98.1** (99.1)	84.0 (86.1)	95.9 (96.3)	90.2 (91.8)
Cosmos2.5	**98.7** (98.8)	98.0 (99.2)	**84.2** (87.0)	**95.9** (96.4)	**90.4** (92.4)
Wan2.2-5B	98.6 (98.5)	98.0 (99.1)	84.0 (86.5)	95.8 (96.1)	90.3 (91.9)

Table 8: Ablation on action expert scaling.

Action Params (B)	NC	DAC	EP	TTC	PDMS	FLOPs (G)↓
0.21	98.6	97.8	84.0	95.4	89.9	27.3
0.45	98.6	97.9	83.8	**95.9**	90.1	55.3
1.02	**98.6**	**98.0**	**84.0**	95.8	**90.3**	112.8

Table 9: Ablation on the co-training signals.

ID	Video Expert Training	Future-Video Supervision	RL	NC	DAC	EP	TTC	PDMS
1	-	-	-	98.0	95.2	81.0	93.8	86.5
2	✓	-	-	98.2	96.8	82.6	94.5	88.4
3	✓	✓	-	**98.6**	98.0	84.0	95.8	90.3
4	✓	✓	✓	98.5	**99.1**	**86.5**	**96.1**	**91.9**

representations at inference. In contrast, our isolated mask lets the action expert attend only to the current observation, so trajectory inference never touches future-frame synthesis. Despite its simpler dependency structure, the isolated mask achieves the best PDMS of 90.3 while predicting the trajectory nearly 3× faster than the masks that condition on future frames. These results suggest that future-video prediction can serve as training supervision rather than a test-time requirement for effective planning.

Analysis on video expert flexibility. SimWAM accommodates diverse pretrained video generators through a unified attention interface, as summarized in Tab. 7. Wan2.1-1.3B and Wan2.2-5B attain comparable PDMS of 90.2 and 90.3 under IL and 91.8 and 91.9 after RL, confirming that our method remains independent of any particular video backbone. Notably, the newer Cosmos-Predict2.5 (Ali et al., 2025), pretrained on driving videos, provides a stronger driving-dynamics prior and achieves the best PDMS of 90.4 under IL and 92.4 after RL. By comparison, the lightweight LTX-Video (HaCohen et al., 2024) reaches 88.7 under IL and 90.9 after RL, indicating that the quality of this driving prior still matters. These results highlight that SimWAM can seamlessly absorb stronger and more driving-relevant priors from advanced video generators.

Analysis on action expert scalability. The parameter-decoupled two-expert design further allows independent scaling of the action expert, as listed in Tab. 8. Under imitation learning, PDMS climbs from 89.9 to 90.3 as the action DiT grows from 0.21B to 1.02B with a fourfold FLOPs increase, and this scaling preserves the video expert and training objective. The width can therefore be traded against the deployment budget. Unless otherwise stated, we use the 1.02B action expert because it achieves the highest PDMS.

Analysis on the co-trained representation. Tab. 9 and Fig. 3 jointly examine how co-training shapes the representation that drives planning. Fine-tuning the video expert using the action loss improves PDMS from 86.5 (ID 1) to 88.4 (ID 2). Adding future-video supervision further increases PDMS to 90.3 (ID 3). This consistent progression indicates that supervision over future frames instills a strong traffic-dynamics prior that directly benefits trajectory planning. Action RL (ID 4) then contributes a further 1.6, demonstrating that optimizing against the driving reward mitigates the conservative predictions of the imitation-trained policy.

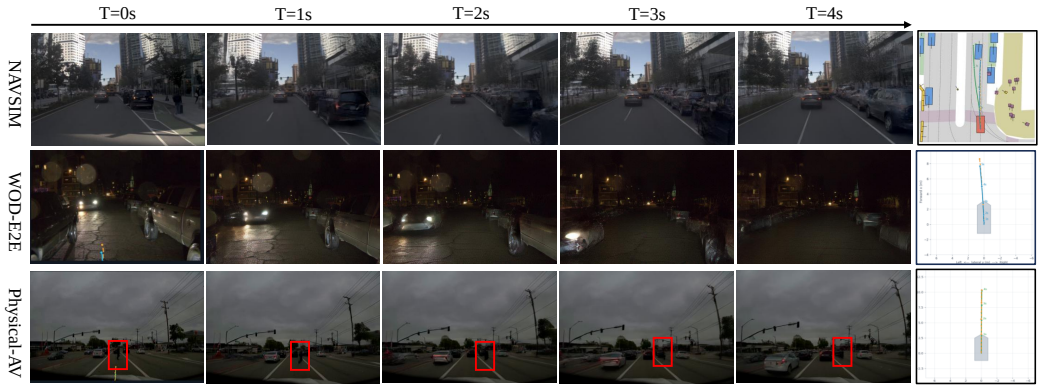


Figure 5: Future-video and trajectory predictions of SimWAM on NAVSIM, WOD-E2E, and PhysicalAI-AV. The red box marks a pedestrian entering the crosswalk.

Beyond planning accuracy, we verify this prior at the representation level. Since SimWAM reads only current-frame features when planning at inference, we probe precisely these features of the video DiT as trained under each configuration above. Two lightweight probes are attached to them. Probe A linearly predicts the motion of surrounding vehicles from ROI-pooled features, whereas Probe B regresses the ego trajectory from globally pooled ones. Both are trained on a 10k-scene subset and evaluated on the test split. Under both probes, the co-trained features (ID 3) attain a markedly lower ADE than those of IDs 1 and 2, indicating that co-training distills future scene dynamics into the current-frame representation and yields features substantially more pertinent to planning.

Ablation on the placement of the SDE steps. Fig. 4 compares where the SDE noise enters the denoising chain. Applying it only to the first two steps lowers PDMS from the imitation baseline of 90.3 to 89.3, whereas restricting it to the last three steps achieves the highest final PDMS of 91.9. Adding early-step noise to this setting offers no consistent improvement, with a final PDMS of 91.6. We argue that subsequent deterministic steps can attenuate early perturbations, thereby limiting trajectory diversity and reward differences. Late-step noise may instead influence the final trajectories more directly, enabling more effective exploration and more informative reward signals. We therefore confine the SDE to the last three steps.

5.4 Qualitative Results

As shown in Fig. 5, we visualize the predicted future frames and the planned trajectories of SimWAM on NAVSIM, WOD-E2E, and PhysicalAI-Autonomous-Vehicles. Remarkably, across all these scenes the model delivers accurate future predictions with feasible trajectories, whether in narrow urban streets, night-time glare, or signalized intersections. Most notably, when a pedestrian enters the crosswalk, SimWAM plans a smooth deceleration to yield to the pedestrian, while the predicted frames anticipate the pedestrian's motion. The generated videos further remain largely temporally coherent and consistent with the planned trajectories.

6 Conclusion

In this paper, we present **SimWAM**, a simple yet flexible world-action model that treats future-video prediction purely as a training-time supervision signal. Through joint flow matching under an isolated attention mask, the lightweight action expert absorbs the traffic-dynamics prior of a pretrained video expert while remaining free of future frames at inference. The decoupled design further allows the video backbone to be replaced and the action expert to be scaled independently, while reinforcement learning aligns trajectory generation with the driving reward beyond imitation. Empirically, SimWAM achieves 91.9 PDMS on NAVSIM with a favorable trade-off between accuracy and latency among world-model-based planners, transfers zero-shot to nuScenes, and similarly attains competitive planning accuracy on WOD-E2E and PhysicalAI-Autonomous-Vehicles. These results suggest that training-time world modeling offers a practical route to efficient and generalizable end-to-end driving.

AI USE STATEMENT

In this work, we used generative AI tools to assist with experimental design, methodological feedback, andqualitative visualization. Additionally, we used generative AI tools for literature review, manuscript editing, and drafting. We did not use generative AI tools to generate synthetic data, develop theoretical or conceptual frameworks, formulate mathematical claims, assist with proofs, propose or refine hypotheses, implement methods, translate content, clean or reformat datasets, conduct qualitative or thematic analysis, or interpret experimental results.

We reviewed all AI-assisted work. We manually verified the suggested experiment designs through independent implementation and evaluation, checked the qualitative visualizations against the corresponding experimental outputs, and reviewed AI-assisted text against our experimental records and cited literature where applicable. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide detailed descriptions of the model architecture, training objectives, and reinforcement learning procedure in Secs. 4.1 and 4.2. The datasets, evaluation metrics, data splits, and preprocessing settings are described in Appendices B.1 and B.2, while the training configurations, optimization settings, sampling parameters, and computational resources are documented in Appendix B.3. Additional ablations and qualitative results are provided in Appendices C.2 and D to further support the analysis and verification of our findings.

REFERENCES

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

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

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Florent Bartoccioni, Elias Ramzi, Victor Besnier, Shashanka Venkataramanan, Tuan-Hung Vu, Yihong Xu, Loick Chambon, Spyros Gidaris, Serkan Odabas, David Hurych, et al. Vavim and vavam: Autonomous driving through video generative modeling. *arXiv preprint arXiv:2502.15672*, 2025.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. pi0: A vision-language-action flow model for general robot control. In *Proc. of Robotics: Science and Systems*, 2025.

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2020.

Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition ADP3 Workshop*, 2021.

Wei Cao, Marcel Hallgarten, Tianyu Li, Daniel Dauner, Xunjiang Gu, Caojun Wang, Yakov Miron, Marco Aiello, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. Pseudo-simulation for autonomous driving. In *Proc. of Conference on Robot Learning*, 2025.

Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Felipe Codevilla, Matthias Müller, Antonio López, Vladlen Koltun, and Alexey Dosovitskiy. End-to-end driving via conditional imitation learning. In *Proc. of the IEEE Int. Conf. on Robotics and Automation*, 2018.

OpenScene Contributors. Openscene: The largest up-to-date 3d occupancy prediction benchmark in autonomous driving. `https://github.com/OpenDriveLab/OpenScene`, 2023.

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, et al. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *Proc. of Advances in Neural Information Processing Systems*, 2024.

Lan Feng, Yang Gao, Eloi Zablocki, Quanyi Li, Wuyang Li, Sichao Liu, Matthieu Cord, and Alexandre Alahi. Rap: 3d rasterization augmented end-to-end planning. In *Proc. of Intl. Conf. on Learning Representations*, 2026.

Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Dingkang Liang, Chong Zhang, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 2025.

Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Hongwei Xie, Bing Wang, Guang Chen, Dingkang Liang, and Xiang Bai. Minddrive: A vision-language-action model for autonomous driving via online reinforcement learning. In *Proc. of European Conference on Computer Vision*, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Nature*, 2025.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *Proc. of Intl. Conf. on Learning Representations*, 2022a.

Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *Proc. of European Conference on Computer Vision*, 2022b.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2023.

Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 2023.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. In *Proc. of Conference on Robot Learning*, 2024.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proc. of Intl. Conf. on Learning Representations*, 2014.

Jingyu Li, Junjie Wu, Dongnan Hu, Xiangkai Huang, Bin Sun, Zhihui Hao, Xianpeng Lang, Xiatian Zhu, and Li Zhang. Sgdrive: Scene-to-goal hierarchical world cognition for autonomous driving. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2026a.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. In *Proc. of Robotics: Science and Systems*, 2026b.

Yingyan Li, Yuqi Wang, Yang Liu, Jiawei He, Lue Fan, and Zhaoxiang Zhang. End-to-end driving with online trajectory evaluation via bev world model. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 2025.

Yingyan Li, Shuyao Shang, Weisong Liu, Bing Zhan, Haochen Wang, Yuqi Wang, Yuntao Chen, Xiaoman Wang, Yasong An, Chufeng Tang, et al. Drivevla-w0: World models amplify data scaling law in autonomous driving. In *Proc. of Intl. Conf. on Learning Representations*, 2026c.

Yongkang Li, Kaixin Xiong, Xiangyu Guo, Fang Li, Sixu Yan, Gangwei Xu, Lijun Zhou, Long Chen, Haiyang Sun, Bing Wang, et al. Recogdrive: A reinforced cognitive framework for end-to-end autonomous driving. In *Proc. of Intl. Conf. on Learning Representations*, 2026d.

Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, et al. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *Proc. of Intl. Conf. on Learning Representations*, 2023.

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 *Proc. of Advances in Neural Information Processing Systems*, 2025.

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

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *Proc. of Intl. Conf. on Learning Representations*, 2023.

Zhaohong Liu, Hao Ye, Xianlin Zhang, and Mengshi Qi. Critiquedrivevlm: From verifier-guided reinforcement learning to latent thought distillation for autonomous driving. *arXiv preprint arXiv:2607.04179*, 2026b.

Zhe Liu, Runhui Huang, Rui Yang, Siming Yan, Zining Wang, Lu Hou, Di Lin, Xiang Bai, and Hengshuang Zhao. Drivepi: Spatial-aware 4d mllm for unified autonomous driving understanding, perception, prediction and planning. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2026c.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. of Intl. Conf. on Learning Representations*, 2019.

Yingzi Ma, Yulong Cao, Wenhao Ding, Shuibai Zhang, Yan Wang, Boris Ivanovic, Ming Jiang, Marco Pavone, and Chaowei Xiao. dvlm-ad: Enhance diffusion vision-language-model for driving via controllable reasoning. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 2023.

Qihang Peng, Xuesong Chen, Chenye Yang, Shaoshuai Shi, and Hongsheng Li. Colavla: Leveraging cognitive latent reasoning for hierarchical parallel trajectory planning in autonomous driving. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2026.

Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. In *Proc. of Advances in Neural Information Processing Systems*, 1988.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 2020.

Ishaan Rawal, Shubh Gupta, Yihan Hu, and Wei Zhan. Nord: A data-efficient vision-language-action model that drives without reasoning. *arXiv preprint arXiv:2602.21172*, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Zihao Sheng, Xin Ye, Jingru Luo, Sikai Chen, and Liu Ren. Explorevla: Dense world modeling and exploration for end-to-end autonomous driving. In *Proc. of European Conference on Computer Vision*, 2026.

Chen Shi, Jinrui Xu, Shaoshuai Shi, Kehua Sheng, Bo Zhang, and Li Jiang. Drivewam: Video generative priors enable scalable world-action modeling for autonomous driving. *arXiv preprint arXiv:2605.28544*, 2026.

Shuhan Tan, Kashyap Chitta, Yuxiao Chen, Ran Tian, Yurong You, Yan Wang, Wenjie Luo, Yulong Cao, Philipp Krähenbühl, Marco Pavone, and Boris Ivanovic. Latent chain-of-thought world modeling for end-to-end autonomous driving. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2026.

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

Daming Wang, Yuhao Song, Zijian He, Kangliang Chen, Xing Pan, Lu Deng, and Weihao Gu. Hmvlm: Multistage reasoning-enhanced vision-language model for long-tailed driving scenarios. *arXiv preprint arXiv:2506.05883*, 2025a.

Hang Wang, Xin Ye, Feng Tao, Chenbin Pan, Abhirup Mallik, Burhan Yaman, Liu Ren, and Junshan Zhang. Adawm: Adaptive world model based planning for autonomous driving. In *Proc. of Intl. Conf. on Learning Representations*, 2025b.

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

Yan Wang, Wenjie Luo, Junjie Bai, Yulong Cao, Tong Che, Ke Chen, Yuxiao Chen, Jenna Diamond, Yifan Ding, Wenhao Ding, et al. Alpamayo-r1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail. *arXiv preprint arXiv:2511.00088*, 2025c.

Tianze Xia, Yongkang Li, Lijun Zhou, Jingfeng Yao, Kaixin Xiong, Haiyang Sun, Bing Wang, Kun Ma, Guang Chen, Hangjun Ye, et al. Drivelaw: Unifying planning and video generation in a latent driving world. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2026.

Runsheng Xu, Hubert Lin, Wonseok Jeon, Hao Feng, Yuliang Zou, Liting Sun, John Gorman, Kate Tolstaya, Sarah Tang, Brandyn White, et al. Wod-e2e: Waymo open dataset for end-to-end driving in challenging long-tail scenarios. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2026.

Zhenjie Yang, Xiaosong Jia, Qifeng Li, Xue Yang, Maoqing Yao, and Junchi Yan. Raw2drive: Reinforcement learning with aligned world models for end-to-end autonomous driving (in carla v2). In *Proc. of Advances in Neural Information Processing Systems*, 2025.

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

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026.

Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, and Xing Wei. Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. In *Proc. of Advances in Neural Information Processing Systems*, 2025.

Bozhou Zhang, Nan Song, jingyu Li, Xiatian Zhu, Jiankang Deng, and Li Zhang. Future-aware end-to-end driving: Bidirectional modeling of trajectory planning and scene evolution. In *Proc. of Advances in Neural Information Processing Systems*, 2025a.

Dongkun Zhang, Jiaming Liang, Ke Guo, Sha Lu, Qi Wang, Rong Xiong, Zhenwei Miao, and Yue Wang. Carplanner: Consistent auto-regressive trajectory planning for large-scale reinforcement learning in autonomous driving. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2025b.

Kaiwen Zhang, Zhenyu Tang, Xiaotao Hu, Xingang Pan, Xiaoyang Guo, Yuan Liu, Jingwei Huang, Li Yuan, Qian Zhang, Xiao-Xiao Long, et al. Epona: Autoregressive diffusion world model for autonomous driving. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 2025c.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2025a.

Zhida Zhao, Talas Fu, Yifan Wang, Lijun Wang, and Huchuan Lu. From forecasting to planning: Policy world model for collaborative state-action prediction. In *Proc. of Advances in Neural Information Processing Systems*, 2025b.

Wenzhao Zheng, Ruiqi Song, Xianda Guo, Chenming Zhang, and Long Chen. Genad: Generative end-to-end autonomous driving. In *Proc. of European Conference on Computer Vision*, 2024.

Xin Zhou, Dingkang Liang, Sifan Tu, Xiwu Chen, Yikang Ding, Dingyuan Zhang, Feiyang Tan, Hengshuang Zhao, and Xiang Bai. Hermes: A unified self-driving world model for simultaneous 3d scene understanding and generation. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 2025.

Zewei Zhou, Tianhui Cai, Seth Zhao, Yun Zhang, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. *Proc. of Advances in Neural Information Processing Systems*, 2026.

A Appendix Overview

The appendix is organized as follows:

- Appendix B: Dataset statistics, evaluation metrics, and training details.
- Appendix C: Additional results on NAVSIM-v2 and further ablations.
- Appendix D: More qualitative results.

B More Details on Datasets and Experimental Setup

B.1 Datasets

NAVSIM is derived from the OpenScene (Contributors, 2023) subset of nuPlan (Caesar et al., 2021). We train on the `navtrain` split with $85{,}109$ samples and evaluate on the `navtest` split with $12{,}146$ samples. On WOD-E2E (Xu et al., 2026) we exclude training samples whose future frames are incomplete, which leaves $333{,}537$ samples, and evaluate on the rater-annotated validation and test splits. PhysicalAI-Autonomous-Vehicles contains roughly $1{,}700$ hours of driving logs organized into $306{,}152$ clips of 20 seconds, with $153{,}625$ clips for training, $90{,}928$ for validation, and $61{,}599$ for testing. We use the front-view camera stream and ego-motion labels. Following DriveWAM (Shi et al., 2026), we use 65K training samples from the training clips for imitation learning and evaluate on the same $1{,}000$-clip test subset for a consistent comparison.

B.2 Evaluation Metrics

NAVSIM PDMS. NAVSIM (Dauner et al., 2024) evaluates each predicted trajectory via non-reactive pseudo-simulation over 4 s, scoring no at-fault collision (NC), drivable area compliance (DAC), ego progress (EP), time-to-collision (TTC) within bound, and comfort (C). The scene score is:

$$\mathrm{PDMS} = \mathrm{NC} \cdot \mathrm{DAC} \cdot \frac{5\,\mathrm{EP} + 5\,\mathrm{TTC} + 2\,\mathrm{C}}{12}. \tag{8}$$

NAVSIM-v2 EPDMS and navhard. NAVSIM-v2 (Cao et al., 2025) adds driving direction compliance (DDC), traffic light compliance (TLC), lane keeping (LK), and extended comfort (EC), and refines C into history comfort (HC). Its scene score is:

$$\mathrm{EPDMS} = \mathrm{NC} \cdot \mathrm{DAC} \cdot \mathrm{DDC} \cdot \mathrm{TLC} \cdot \frac{5\,\mathrm{EP} + 5\,\mathrm{TTC} + 2\,\mathrm{LK} + 2\,\mathrm{HC} + 2\,\mathrm{EC}}{16}, \tag{9}$$

where a term scores one whenever its human counterpart scores zero. NAVSIM-v2 further introduces `navhard`, a safety-critical subset evaluated in two stages. The first stage scores the planner on real-world scenarios, and the second re-evaluates the corresponding synthesized scenarios with traffic agents reacting to the ego vehicle.

WOD-E2E RFS. WOD-E2E (Xu et al., 2026) compares a prediction with three human-rated references p, each scored $s_p \in [0, 10]$. At $t \in \{3, 5\}$ s, let $d_{p,t}$ be the larger axis-wise displacement between the prediction and reference p, normalized by the speed-scaled trust-region thresholds. The per-reference and final scores are:

$$\begin{aligned} q_{p,t} &= s_p\, 0.1^{\max(d_{p,t}-1,0)}, \qquad S = \frac{1}{2} \sum_{t\in\{3,5\}} \max_p q_{p,t}, \\ \mathrm{RFS} &= \begin{cases} S, & \text{if } \max_{t\in\{3,5\}} d_{p,t} \le 1 \text{ for some } p, \\ \max(S, 4), & \text{otherwise.} \end{cases} \end{aligned} \tag{10}$$

The official test RFS averages scenario-cluster scores, with the lower bound applied after horizon averaging unless both endpoints lie within one reference's trust region. ADE@3s/@5s averages position errors at 4 Hz through each horizon.

Table 10: Performance on the NAVSIM-v2 `navtest` benchmark. $*$ indicates training with reinforcement learning. The best learned result in each column is shown in **bold**.

Method	Reference	NC↑	DAC↑	DDC↑	TLC↑	EP↑	TTC↑	LK↑	HC↑	EC↑	EPDMS↑
Human Agent	-	100.0	100.0	99.8	100.0	87.4	100.0	100.0	98.1	90.1	90.3
Traditional E2E planners											
TransFuser (Chitta et al., 2022)	TPAMI'22	96.9	89.9	97.8	99.7	87.1	95.4	92.7	98.3	87.2	76.7
DiffusionDrive (Liao et al., 2025)	CVPR'25	98.2	95.9	99.4	99.8	87.5	97.3	96.8	98.3	**87.7**	84.5
VLM-based planners											
ReCogDrive$*$ (Li et al., 2026d)	ICLR'26	98.3	95.2	99.5	99.8	87.1	97.5	96.6	98.3	86.5	83.6
SGDrive (Li et al., 2026a)	CVPR'26	98.6	94.3	99.5	**99.9**	86.0	97.9	96.1	98.3	85.9	86.2
World-model-based planners											
DriveVLA-W0 (Li et al., 2026c)	ICLR'26	98.5	**99.1**	98.0	99.7	86.4	98.1	93.2	97.9	58.9	86.1
DriveLaW (Xia et al., 2026)	CVPR'26	**98.7**	96.9	99.6	99.8	87.5	98.3	97.6	**98.4**	77.4	88.6
SimWAM (**ours**)	-	98.6	98.0	**99.7**	**99.9**	87.5	**98.4**	**97.9**	98.3	84.4	**90.2**

Table 11: Performance on the NAVSIM-v2 `navhard` benchmark. S1/S2 denote the per-stage results of the official two-stage protocol, and EPDMS is the overall score. The best learned result in each column is shown in **bold**.

Method	Reference	Stage	NC↑	DAC↑	DDC↑	TLC↑	EP↑	TTC↑	LK↑	HC↑	EC↑	EPDMS↑
Traditional E2E planners												
TransFuser (Chitta et al., 2022)	TPAMI'22	S1	96.2	79.5	99.1	99.5	84.1	95.1	94.2	97.5	79.1	23.1
		S2	77.7	70.2	84.2	98.0	85.1	75.6	45.4	95.7	**75.9**	
DiffusionDrive (Liao et al., 2025)	CVPR'25	S1	96.8	86.0	98.8	99.3	84.0	95.8	96.7	97.6	**79.6**	27.5
		S2	80.1	72.8	84.4	98.4	85.9	76.6	46.4	96.3	72.8	
VLM-based planners												
SGDrive (Li et al., 2026a)	CVPR'26	S1	95.8	87.6	97.8	**99.8**	84.4	94.7	92.9	**97.8**	28.9	25.5
		S2	79.4	65.4	79.1	**98.9**	88.9	75.3	42.7	96.4	29.6	
ReCogDrive (Li et al., 2026d)	ICLR'26	S1	96.4	78.9	98.7	**99.8**	82.6	95.6	94.4	97.6	74.2	25.7
		S2	80.2	65.0	82.4	98.7	85.2	76.9	43.8	**96.6**	71.8	
World-model-based planners												
DriveVLA-W0 (Li et al., 2026c)	ICLR'26	S1	96.8	83.3	99.0	99.6	**84.6**	95.3	96.4	97.6	78.2	24.4
		S2	76.8	64.3	79.9	98.3	**89.2**	75.0	46.8	95.8	53.1	
DriveLaW (Xia et al., 2026)	CVPR'26	S1	97.3	89.1	99.2	99.6	84.3	**97.1**	96.2	**97.8**	67.6	30.6
		S2	**82.5**	67.6	83.5	98.1	84.8	78.5	45.8	96.4	57.3	
SimWAM (**ours**)		S1	**98.0**	**92.0**	**99.7**	99.6	83.8	96.2	**97.3**	**97.8**	71.6	**37.6**
		S2	81.8	**78.6**	**87.3**	98.4	86.3	**78.6**	**49.5**	96.3	69.9	

B.3 Training Details

On NAVSIM the action expert predicts 8 waypoints over 4 s at 2 Hz while the video expert predicts the corresponding 8 future frames. On WOD-E2E we predict 20 waypoints over 5 s at 4 Hz, and on PhysicalAI we report the ADE and FDE over 3 s and 4 s horizons. On both WOD-E2E and PhysicalAI the video expert predicts 4 future frames over 4 s at 1 Hz. All imitation-learning runs adopt AdamW (Loshchilov & Hutter, 2019) and a cosine learning rate schedule with an initial learning rate of 10^{-4}, with λ=1 across the three datasets. During reinforcement learning (RL), we optimize rank-16 LoRA adapters (Hu et al., 2022a) with scale α=32 on the attention projections of the action expert. We sample G=8 trajectories per scenario with noise level η=0.1 restricted to the last 3 of the 10 sampling steps, and use a learning rate of 5×10^{-5}, an anchor weight of β=50, coordinate weights w=$(0.2, 1, 1)$, and ζ=10^{-8}. On NAVSIM, RL trains for $20{,}000$ steps on the `navtrain` scenes whose imitation PDMS falls below 90, while evaluation covers the full `navtest` split. On WOD-E2E, RL optimizes the RFS reward for 10 epochs on the validation split. Imitation learning runs on 48 A100 GPUs with total batch size 192, while RL runs on a single A100 GPU.

C Additional Experimental Results

C.1 NAVSIM-v2 Results

We also evaluate SimWAM under NAVSIM-v2 (Cao et al., 2025) on both `navtest` and the safety-critical `navhard` benchmark, as shown in Tabs. 10 and 11. Even before reinforcement learning, SimWAM attains the highest EPDMS of 90.2 on `navtest` and 37.6 on `navhard` among the compared methods, surpassing DriveLaW (Xia et al., 2026) by 1.6 EPDMS points on `navtest` and by 7.0 on `navhard`, with leading DAC, DDC, TTC, and LK, especially in the reactive second stage

Table 12: Ablation on future-video supervision.

Setting	NC	DAC	EP	TTC	PDMS
4 f, 2 s, 2 Hz	98.6	97.7	83.9	95.5	89.9
4 f, 4 s, 1 Hz	**98.7**	97.9	**84.2**	95.6	90.2
8 f, 4 s, 2 Hz	98.6	**98.0**	84.0	**95.8**	**90.3**

Table 13: The effect of input resolution.

Resolution	NC	DAC	EP	TTC	PDMS	Latency (ms)↓
192×352	98.2	97.1	83.0	94.9	88.9	509
384×672	98.6	98.0	84.0	95.8	90.3	518
768×1344	**98.7**	**98.1**	**84.3**	**96.1**	**90.6**	573

Table 14: The effect of sampling steps.

Steps	NC	DAC	EP	TTC	PDMS	Latency (ms)↓
1	97.4	91.3	79.1	83.3	68.9	115
5	**98.6**	97.9	**84.0**	95.6	90.1	297
10	**98.6**	**98.0**	**84.0**	**95.8**	**90.3**	518
20	**98.6**	**98.0**	83.9	**95.8**	90.2	968

Table 15: Ablation of the RL scheme.

Anchor Loss	w	NC	DAC	EP	TTC	PDMS
-	0.2,1,1	94.7	94.4	85.1	87.5	84.3
✓	1,1,1	98.5	99.0	85.5	96.2	91.6
✓	0.5,1,1	**98.6**	99.0	85.7	**96.4**	91.8
✓	0.2,1,1	98.5	**99.1**	**86.5**	96.1	**91.9**

where surrounding agents respond to the ego vehicle. Collectively, these results show that training-time world modeling delivers superior planning without any future-frame synthesis at inference. Moreover, the learned dynamics prior remains effective even when traffic reacts to the planner.

C.2 Additional Ablations

Ablation on the prediction horizon and frame density. We examine the temporal configuration of future-video supervision in Tab. 12. Shortening the prediction horizon from 4 s to 2 s noticeably reduces PDMS, whereas maintaining the 4 s horizon with half as many frames recovers most of the performance. This comparison indicates that broad temporal coverage is more important than dense frame sampling for learning traffic dynamics. The full 4 s target at 2 Hz achieves the strongest performance.

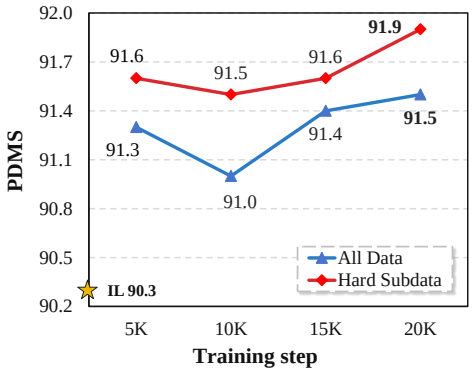


Figure 6: Ablation on the RL training data. The challenging subset with imitation PDMS below 90 consistently outperforms the full `navtrain` set.

Ablation on the input resolution. We next study the trade-off between visual detail and inference efficiency in Tab. 13. Increasing the resolution from 192×352 to 384×672 substantially improves PDMS by 1.4 points with only 9 ms of additional latency. Further increasing the resolution to 768×1344 yields merely a 0.3 point gain while adding considerably more computation. These results identify 384×672 as the most favorable balance between planning accuracy and inference efficiency.

Ablation on the number of sampling steps. We further investigate the convergence of the action flow sampler in Tab. 14. A single sampling step is insufficient to produce well-refined trajectories, whereas five steps already recover most of the performance. Increasing the budget to ten steps achieves the highest PDMS of 90.3. Twenty steps provide no further improvement while nearly doubling the latency, indicating that performance saturates around ten steps in our setting.

Ablation on the reinforcement-learning scheme. Tab. 15 ablates the two design choices of our reinforcement-learning stage under a common training budget. Raising the longitudinal weight of the mean anchor to 0.5 or 1.0 slightly reduces PDMS, confirming that the *down-weighted longitudinal coordinate* leaves the policy more freedom to optimize the driving reward. In contrast, removing the anchor loss collapses PDMS to 84.3, even below the imitation checkpoint, showing that the anchor is *essential* to stable policy updates rather than an optional regularizer.

RL training data. Fig. 6 compares reinforcement learning on the full `navtrain` split and on the challenging subset whose imitation-policy PDMS is below 90. The subset consistently outperforms the full set throughout training and reaches 91.9 at 20k steps. The difficult scenarios expose clearer differences among the sampled trajectories and thus provide more informative reward signals. In contrast, many scenes in the full set are already well handled by imitation learning and contribute limited learning signals, which motivates our subset restriction.



Figure 7: Additional qualitative results of SimWAM across NAVSIM, WOD-E2E, and PhysicalAI-AV, covering diverse streets, intersection turns, large leading trucks, and snow-covered roads.

D More Qualitative Results

D.1 Future-Video and Trajectory Visualization

As shown in Fig. 7, we present qualitative results on more diverse scenarios, spanning dense urban streets, left and right turns at intersections, close following of large trucks, and snow-covered roads across the three benchmarks. In each case, SimWAM anticipates the evolution of the surrounding traffic over the four-second horizon, and the predicted video stream remains consistent with the planned trajectory shown in the right column. Notably, the snow-covered example presents challenging visual conditions, while the predicted trajectory remains visually smooth. These examples corroborate the coherent and reliable behavior of SimWAM under varied driving conditions.

D.2 Trajectory Visualization before and after Reinforcement Learning

Fig. 8 compares the imitation-trained and RL-fine-tuned models on two scenes. The imitation-trained model produces conservative trajectories and advances only a short distance at the intersection and along the narrow street. After reinforcement learning, the model follows the intended route more decisively and completes a larger portion of each maneuver, while the trajectories remain within the drivable area and keep a safe clearance from surrounding vehicles. This comparison shows that the driving reward mitigates the conservative predictions of the imitation-trained policy.



Concat View Ours-IL Ours-RL

Figure 8: Qualitative comparison of *Ours-IL* and *Ours-RL* on two `navtest` scenarios. Red ellipses highlight regions where *Ours-RL* progresses farther while remaining within the drivable area.