TRAJWAM: TRAJECTORY-CONDITIONED WORLD ACTION MODELING FOR VISION–LANGUAGE DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

World Action Models (WAMs) connect future scene prediction with action planning for autonomous driving. Multi-view video prediction requires camera-wise image decoding, while trajectory-agnostic bird's-eye-view (BEV) prediction leaves the influence of future ego motion implicit. Thus, we present TrajWAM, a trajectory-conditioned WAM that jointly learns BEV future representations and a vision-language driving policy. TrajWAM links world prediction and planning through a shared backbone by conditioning DINOv3-aligned BEV predictions on ego trajectories. Training combines reconstruction of observed futures with a feature-separation objective that encourages predictions to respond to trajectory changes, without requiring future labels for perturbed trajectories. In inference, the planner directly generates waypoints while bypassing the auxiliary future-prediction branch. On all 220 Bench2Drive routes, TrajWAM achieves 91.53 Driving Score and 75.45% Success Rate, exceeding a schedule-matched planning-only fine-tuning baseline by 5.31 Driving Score points and eight successful routes.

1 INTRODUCTION

Autonomous driving requires recognizing the current scene and anticipating how it may evolve under ego actions. World Action Models (WAMs) couple observations and action queries with representations of what may follow (Wang et al., 2026d; Ye et al., 2026). Driving systems learn video, latent, or bird's-eye-view (BEV) futures to support planning (Wang et al., 2024; Li et al., 2025a;c; Zheng et al., 2025). Such prediction can also provide training supervision when photorealistic synthesis is not a deployment objective (Wang et al., 2026b).

Diffusion-based video prediction entails decoding multiple camera streams and future frames. Latent and BEV prediction already offer alternatives to this image-space interface, including future supervision within driving VLMs (Zhang et al., 2026) and auxiliary latent dynamics that are bypassed at inference (Wang et al., 2026c). The first comparison in Figure 1 motivates this representation-space choice through a conceptual decoding-cost comparison.

However, predicting future BEV representations alongside driving trajectories does not by itself make future prediction explicitly trajectory-conditioned. In a trajectory-agnostic VLM, future reconstruction supervises what follows the observed scene without taking a future ego trajectory as an explicit query. We introduce this query into the VLM's future-representation learning so that the predicted future depends on both the scene and the supplied trajectory. The second comparison in Figure 1 illustrates this distinction.

Trajectory conditioning raises a supervision challenge: each training record provides one executed trajectory and its observed future, leaving alternative trajectories without observed outcomes. Although a model can learn trajectory dependencies across records, each record directly supervises only its logged query. This motivates our central question: *How can trajectory-conditioned BEV future learning improve a VLM planner when only the logged trajectory has an observed future?*

We introduce TrajWAM, a *trajectory-conditioned World Action Model* built on a Qwen2.5-VL-3B planner (Bai et al., 2025). Its world component predicts DINOv3-aligned BEV representations, its action query is a four-waypoint ego trajectory, and future prediction and waypoint generation train a

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.

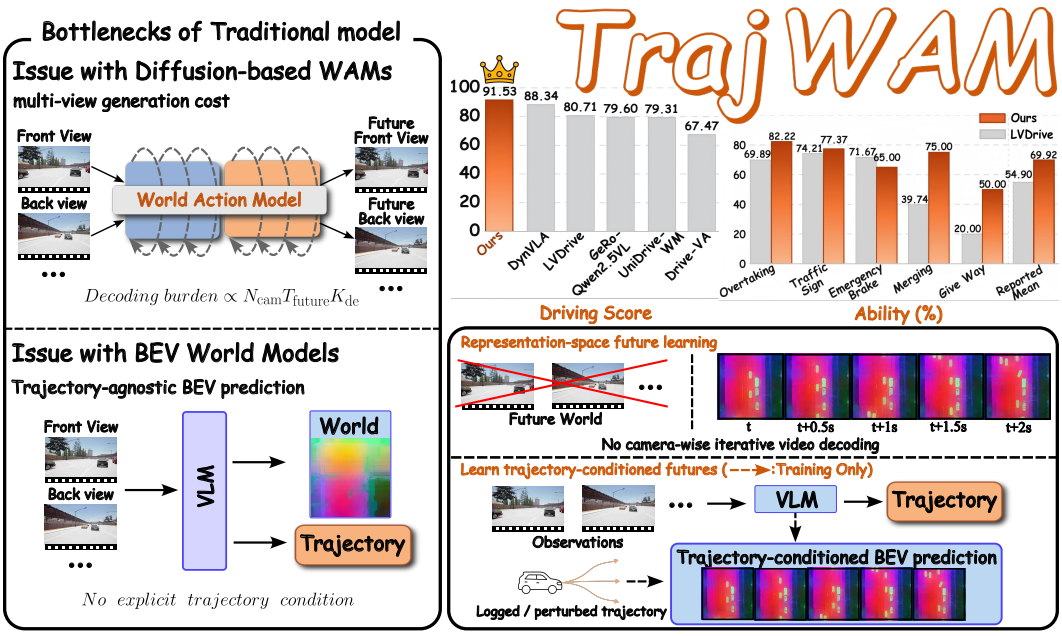

Figure 1: Motivation and overview. The comparisons illustrate representation-space prediction without camera-wise decoding and trajectory-conditioned supervision beyond the illustrated factual-only BEV interface. The decoding-cost comparison is conceptual. Only the logged query has future targets; the perturbed query receives a separation loss, and the readout is bypassed at inference. Feature maps use AnyUp (Wimmer et al., 2026) and PCA for visualization; details are provided in Appendix A.

shared VLM backbone. The supplied trajectory modulates future-query states through feature-wise linear modulation (FiLM) (Perez et al., 2018); a readout predicts one current anchor and four future representations (Siméoni et al., 2025). This world–action coupling operates during training. Native inference retains the fixed future-query tokens but bypasses trajectory conditioning and the DINO readout, generating reasoning and waypoints without future rollout or candidate reranking.

The objective distinguishes queries by their available supervision. The logged trajectory $A_t^\star$ is paired with the observed DINOv3 target sequence. Its coordinate-wise perturbation A_t^- receives no future target and participates only in a one-sided feature-separation loss. This objective encourages the predicted representation to respond to changes in the trajectory query. Both trajectories enter the auxiliary readout after planner-logit computation; their losses influence planning through shared training gradients.

We evaluate TrajWAM on all 220 Bench2Drive routes (Jia et al., 2024). A schedule-matched continuation compares the full method with additional planning-only fine-tuning, and component ablations compare factual reconstruction, trajectory conditioning, and separation. Figure 1 previews the performance, with detailed comparisons in Section 4.

Our contributions are:

1. **Trajectory-conditioned world–action learning within a VLM.** An external four-waypoint query conditions planner-native future states through FiLM, coupling DINOv3-aligned BEV prediction with waypoint generation through shared training gradients.
2. **A supervision objective for single-outcome logs.** Factual reconstruction supervises the prediction under the logged query, while a one-sided objective encourages separation from the prediction under a perturbed query without assigning it an unobserved future target.
3. **Closed-loop evaluation of the training design.** On all 220 Bench2Drive routes, the full objective achieves higher DS and SR than a schedule-matched planning-only continuation and the ablated variants, with the auxiliary readout bypassed at native inference.

2 Related Work

World action models and video-generative priors. DreamZero (Ye et al., 2026) jointly models video and robot actions with a pretrained video-diffusion backbone. Drive-WM (Wang et al., 2024) synthesizes maneuver-conditioned multi-view driving videos for trajectory selection. GenCeption (Wang et al., 2026b) studies transferable visual representations and vision–language alignment from video-generative pretraining.

Representation-space futures in driving. LAW (Li et al., 2025a) predicts perspective and BEV latents conditioned on ego trajectories. WoTE (Li et al., 2025c) and World4Drive (Zheng et al., 2025) use candidate-dependent futures for online trajectory selection. UniDrive-WM (Xiong et al., 2026), DriveWorld-VLA (Jia et al., 2026), VLA-World (Wang et al., 2026a), and Latent-CoT-Drive (Tan et al., 2026) integrate future generation or latent imagination with VLM/VLA planning. DriveVLA-W0 (Li et al., 2025b) makes future-image prediction optional at inference, while HyWorldVLA (Yu et al., 2026) combines latent prediction with pixel reconstruction.

Planner-native semantic futures. DeepSight (Zhang et al., 2026) predicts DINOv3-aligned BEV futures alongside VLM reasoning and planning. Latent-WAM (Wang et al., 2026c) predicts latent transitions from scene and ego-state history, including commands, velocity, and acceleration, using EMA-encoder targets and omitting its dynamics module at inference. TrajWAM instead conditions VLM future-query states on an external future-waypoint trajectory and reconstructs fixed DINOv3 targets for the logged query. It extends DeepSight-style factual prediction with explicit trajectory conditioning and perturbed-query separation. DINOv3-aligned BEV supervision is shared with DeepSight; omitting auxiliary dynamics at inference is also used by Latent-WAM.

Training-time supervision and candidate evaluation. EgoWAM (Li et al., 2026) compares pixel-, DINO-, and motion-targeted supervision. Fast-WAM (Yuan et al., 2026) and SimWAM (Zhao et al., 2026b) bypass future prediction at inference, whereas Faster-WAM (Zhao et al., 2026a) retains sparse future conditioning. DA-WAM (Zhong et al., 2026) addresses single-outcome driving logs by supervising the expert-matched candidate with future latents and all candidates with factor, utility, and ranking targets for online evaluation. In TrajWAM, perturbed queries receive only a feature-separation constraint; the native policy performs no candidate scoring.

3 Method

3.1 Problem Setup and Base Planner

TrajWAM couples an action-generating VLM with an auxiliary, trajectory-conditioned world readout. The two components share planner states and are trained jointly; the external action query affects future-representation prediction, while its supervision shapes waypoint generation through shared gradients.

At time t, the planner receives $X_t = (O_t, s_t, r_t, q^{\mathrm{fw}})$, where O_t contains four historical front-camera frames and a current six-camera surround-view observation, s_t is the ego state, r_t is the route/target instruction, and q^{fw} is a fixed block of future-world query tokens. These tokens remain in the native VLM sequence during both training and inference; only their auxiliary trajectory-conditioned readout is training-only. A Qwen2.5-VL-3B planner P_ϕ produces hidden states

$$H_t = P_\phi(X_t) \in \mathbb{R}^{L \times D}, \qquad D = 2048, \tag{1}$$

and autoregressively generates four future waypoints $\widehat{A}_t = \{(\widehat{a}_{k,1}, \widehat{a}_{k,2})\}_{k=1}^{4}$. The inherited PID controller converts these waypoints into low-level control.

We first train the Base Planner (Phase I) with token-level reasoning-and-waypoint supervision and a trajectory-agnostic DINOv3-aligned BEV future-representation objective inspired by DeepSight (Zhang et al., 2026). Starting from the Base Planner (Phase I) checkpoint, TrajWAM introduces the trajectory address and jointly optimizes native waypoint generation with trajectory-conditioned DINOv3-aligned BEV future-representation learning. Let

$$\mathcal{T} = \{0, 0.5, 1.0, 1.5, 2.0\}\ \mathrm{s}, \tag{2}$$

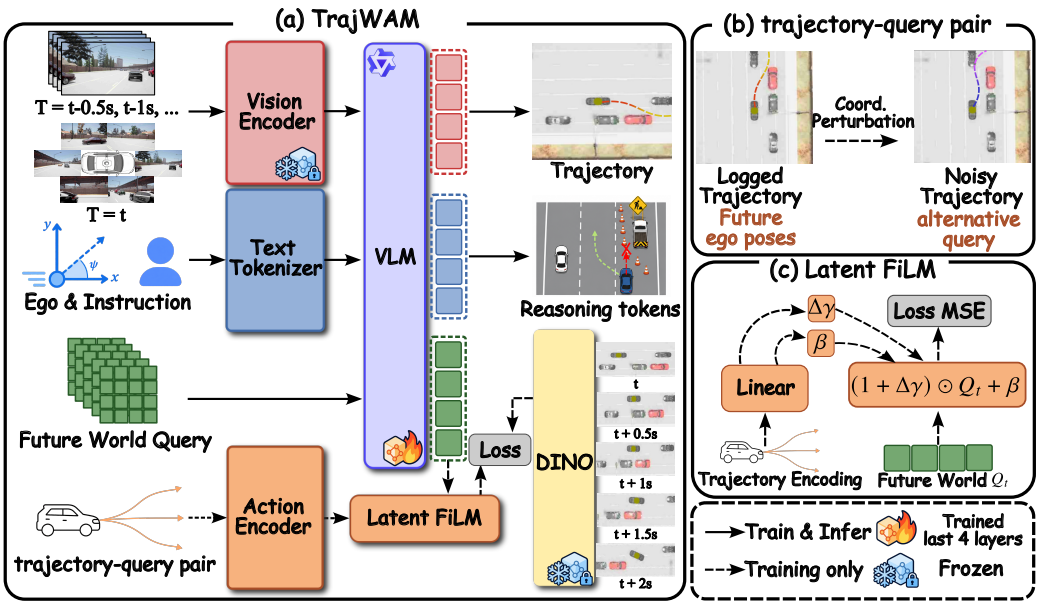

Figure 2: Architecture of TrajWAM. (a) The VLM generates reasoning and four waypoints; during training, an external trajectory conditions its DINOv3 readout through the ActionEncoder and latent FiLM. A frozen teacher supplies logged BEV targets. (b) Coordinate perturbation creates an unlabeled query A_t^-. (c) FiLM starts at the identity. Solid/dashed arrows indicate native/training-only computation; flame/snowflake icons indicate trainable/frozen modules.

denote the five supervised BEV times, comprising the current anchor and four future observations aligned to the current ego frame. The aligned images $\{B^\star_{t+\tau}\}_{\tau\in\mathcal{T}}$ are used only to construct frozen DINOv3 targets; they are never planner inputs.

Figure 2(a) gives the full information flow. The external trajectory is consumed only by the auxiliary readout after planner-logit computation, and the future target appears only in the loss. Neither can enter or alter the same-pass waypoint computation, although their gradients shape the shared planner across training updates.

3.2 Logged and Noisy Trajectory Queries

Each training record supplies one realized ego trajectory $A_t^\star = [a^\star_{t,1}, \dots, a^\star_{t,4}] \in \mathbb{R}^{4\times 2}$, expressed as four future ego waypoints in the current ego coordinate frame. It is a logged label, not the trajectory predicted by P_ϕ. As illustrated in Figure 2(b), we construct a noisy alternative

$$\begin{aligned}(A_t^-)_{k,1} &= (A_t^\star)_{k,1} + \epsilon_k, \quad \epsilon_k \sim \mathcal{N}(0, 2^2),\\ (A_t^-)_{k,2} &= \rho_k (A_t^\star)_{k,2}, \quad \rho_k \sim \mathcal{U}(0.7, 1.3),\end{aligned} \qquad k \in \{1, \dots, 4\}. \tag{3}$$

We apply this operation in the stored-coordinate order shown above. The term *noisy trajectory query* emphasizes its construction; A_t^- is neither executed nor assigned a future-world label.

The pair $\{A_t^\star, A_t^-\}$ is therefore not a set of online trajectory candidates. It is a training interface for asking whether the predicted representation changes for the same observed scene when the trajectory query changes. Only $A_t^\star$ has an observed outcome; the role of A_t^- is restricted to feature-space discrimination under single-outcome supervision.

3.3 Trajectory Address and Latent FiLM

For either query $A \in \{A_t^\star, A_t^-\}$, the ActionEncoder flattens and normalizes its coordinates and maps them to the VLM hidden dimension:

$$\begin{aligned}z_t(A) &= \mathrm{vec}(A)/\sigma, \qquad \sigma = 10,\\ e_t(A) &= W_2\,\mathrm{GELU}(W_1 z_t(A) + b_1) + b_2 \in \mathbb{R}^D.\end{aligned} \tag{4}$$

Let $M^{\mathrm{fw}}_{t,\tau}$ select the valid future-world query tokens at horizon τ and let $Q_{t,\tau} = H_t[M^{\mathrm{fw}}_{t,\tau}]$. The same sample-level address is broadcast over all valid tokens and all five horizons. Latent FiLM produces

$$\begin{aligned}[\Delta\gamma_t(A), \beta_t(A)] &= F_{\mathrm{lat}}(e_t(A)),\\ \widetilde{Q}_{t,\tau}(A) &= (1 + \Delta\gamma_t(A)) \odot Q_{t,\tau} + \beta_t(A).\end{aligned} \tag{5}$$

Zero-initializing the FiLM projection gives the identity transform in Figure 2(c). The trajectory conditions only the auxiliary readout, after planner-logit computation.

3.4 DINOv3-Aligned Prediction from Single-Outcome Logs

A DINO-space projection $V : \mathbb{R}^D \to \mathbb{R}^{1024}$, learned with the Base Planner (Phase I) and frozen during TrajWAM training, maps the modulated states to DINOv3 feature space:

$$\begin{aligned}\widehat{Y}_{t,\tau}(A) &= V\left(\widetilde{Q}_{t,\tau}(A)\right),\\ \widehat{\mathbf{Y}}_t(A) &= \{\widehat{Y}_{t,\tau}(A)\}_{\tau\in\mathcal{T}}.\end{aligned} \tag{6}$$

A frozen DINOv3 ViT-L/16 teacher constructs the factual targets

$$Y^\star_{t,\tau} = \mathrm{sg}\big[\mathrm{DINOv3}\big(B^\star_{t+\tau}\big)\big]_\Omega, \qquad \tau \in \mathcal{T}, \tag{7}$$

where Ω selects the CLS token and valid patch tokens; register tokens are excluded. The factual objective is

$$\mathcal{L}_{\mathrm{fact}} = \mathrm{MSE}_{\tau,\Omega}\left(\widehat{Y}_{t,\tau}(A^\star_t), Y^\star_{t,\tau}\right). \tag{8}$$

No target is attached to $\widehat{Y}_{t,\tau}(A^-_t)$. Instead, we define the valid-token feature distance

$$d_{\mathrm{sem}}(U, Z) = \mathrm{mean}_{\tau,\Omega,c}\,|U_{\tau,\Omega,c} - Z_{\tau,\Omega,c}|, \tag{9}$$

where $c \in \{1, \ldots, 1024\}$ indexes the raw, unnormalized DINO feature channels. We then use the one-sided separation objective

$$\mathcal{L}_{\mathrm{sep}} = \left[m_{\mathrm{sem}} - d_{\mathrm{sem}}\left(\mathrm{sg}\left(\widehat{\mathbf{Y}}_t(A^\star_t)\right), \widehat{\mathbf{Y}}_t(A^-_t)\right)\right]_+, \tag{10}$$

where $[x]_+ = \max(x, 0)$. The stop-gradient blocks the direct gradient through the factual prediction in this loss. Shared-parameter updates through the perturbed-query branch may still change both predictions, while factual reconstruction anchors the logged-query prediction. Equation 10 encourages trajectory-query sensitivity without requiring a future target for A^-_t.

3.5 Objective, Gradient Routing, and Native Inference

The TrajWAM training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{rec}} + 2\mathcal{L}_{\mathrm{fact}} + \lambda_{\mathrm{sep}}\mathcal{L}_{\mathrm{sep}}, \tag{11}$$

where $\mathcal{L}_{\mathrm{rec}}$ is token-level cross-entropy for autoregressive reasoning and waypoint generation, m_{sem} is the DINO-space margin, and λ_{sep} is its loss weight. The factual prediction is detached only on the positive side of $\mathcal{L}_{\mathrm{sep}}$; it remains fully differentiable under $\mathcal{L}_{\mathrm{fact}}$. Within the DINO-only auxiliary branch, the negative prediction and the planner states feeding it are not detached. Thus, $\mathcal{L}_{\mathrm{fact}}$ and $\mathcal{L}_{\mathrm{sep}}$ update the ActionEncoder, latent FiLM, and the final four trainable planner layers through their respective paths; $\mathcal{L}_{\mathrm{rec}}$ updates the planner but not the task-specific modules. The vision tower, multimodal projector, and DINO-space projection remain frozen; the DINO-space projection still transmits auxiliary gradients to its input states. The frozen DINOv3 teacher constructs targets only under stop-gradient. At native closed-loop inference, the fixed future-world query tokens remain in the VLM prompt, which autoregressively generates reasoning and four waypoints for the inherited PID controller. No logged or noisy trajectory, future target, ActionEncoder, latent FiLM, DINO-space projection, or DINOv3 teacher is invoked; the policy performs neither a future rollout nor online trajectory scoring or reranking. Checkpoint storage is reported in Appendix D.

Table 1: Bench2Drive Base Set closed-loop results. Venue reports conference/preprint and year; * denotes expert-feature distillation. Bold/underlined values mark the best/second-best listed results.

Method	Venue	Paradigm	Expert	Closed-loop			
				DS↑	SR(%)↑	Efficiency↑	Smoothness↑
TCP* (Wu et al., 2022)	NeurIPS '22	E2E	Think2Drive	40.70	15.00	54.26	<u>47.80</u>
ThinkTwice* (Jia et al., 2023b)	CVPR '23	E2E	Think2Drive	62.44	31.23	69.33	16.22
DriveAdapter* (Jia et al., 2023a)	ICCV '23	E2E	Think2Drive	64.22	33.08	70.22	16.01
VAD (Jiang et al., 2023)	ICCV '23	E2E	Think2Drive	42.35	15.00	157.94	46.01
GenAD (Zheng et al., 2024)	ECCV '24	E2E	Think2Drive	44.81	15.90	-	-
MomAD (Song et al., 2025)	CVPR '25	E2E	Think2Drive	44.54	16.71	170.21	**48.63**
DriveTransformer (Jia et al., 2025)	ICLR '25	E2E	Think2Drive	63.46	35.01	100.64	20.78
DiffusionDrive (Liao et al., 2025)	CVPR '25	E2E	PDM-Lite	77.68	52.72	-	-
SparseDriveV2 (Sun et al., 2026)	ECCV '26	E2E	Think2Drive	<u>89.15</u>	70.00	199.84	18.32
ReasonPlan (Liu et al., 2025)	CoRL '25	VLM	Think2Drive	64.01	34.55	180.64	25.63
ORION (Fu et al., 2025)	ICCV '25	VLM	Think2Drive	77.74	54.62	151.48	17.38
SimLingo (Renz et al., 2025)	CVPR '25	VLM	PDM-Lite	85.94	66.82	**244.18**	25.49
AutoVLA (Zhou et al., 2026)	NeurIPS '25	VLM	Think2Drive	78.84	57.73	146.93	39.33
DeepSight (Zhang et al., 2026)	ICML '26	VLM	Think2Drive	86.23	71.36	<u>201.71</u>	16.11
UniDrive-WM (Xiong et al., 2026)	ECCV '26	WM	Think2Drive	79.31	56.42	158.65	27.93
GeRo-Qwen2.5VL (Yasarla et al., 2026)	CVPR '26	WM	Think2Drive	79.60	57.80	159.66	44.75
DynVLA (Shang et al., 2026)	ICML '26	WM	Think2Drive	88.34	<u>72.73</u>	-	-
LVDrive (Mei et al., 2026)	arXiv '26	WAM	Think2Drive	80.71	58.26	155.77	14.34
TrajWAM (Ours)	This work	WAM	Think2Drive	**91.53**	**75.45**	171.41	15.27

4 RESULTS

4.1 EXPERIMENTAL SETUP

Phase I specializes Qwen2.5-VL-3B with autoregressive reasoning/waypoint supervision and trajectory-agnostic DINOv3 future reconstruction, following the DeepSight-style recipe but producing our own Base Planner (Phase I) checkpoint. It contains no ActionEncoder, trajectory FiLM, or separation loss.

Phase II starts from this checkpoint and trains TrajWAM for 24 epochs on 226,656 Bench2Drive examples. We use eight A100 GPUs, BF16, an effective batch size of 64, AdamW with cosine decay and 10% warm-up, and a random seed of 42. Learning rates are 2×10^{-5} for the last four planner layers and 2×10^{-4} for TrajWAM modules; no validation-based checkpoint selection is used. Further optimization and parameter details appear in Appendix A.

Model variants. We distinguish the published DeepSight results from our in-house Base Planner (Phase I). All Phase-II variants independently start from this same checkpoint and are fine-tuned (FT) for 24 epochs. Base Planner (planning-only FT) uses only the reasoning-and-waypoint objective during Phase II. Base Planner (unconditioned future FT) also reconstructs factual DINOv3 targets, without an external trajectory query or FiLM. TrajWAM w/o separation uses trajectory-conditioned factual reconstruction, and TrajWAM additionally uses the separation loss. All variants retain reasoning-and-waypoint supervision and share the Phase-I DINOv3 training history; "planning-only" describes the Phase-II objective only.

Closed-loop evaluation uses all 220 routes in the Bench2Drive Base Set, including failed routes, and reports the official Driving Score (DS), Route Completion (RC), and strict Success Rate (SR). Each of our model variants is evaluated in one complete simulator sweep; results for prior methods are taken from their publications.

4.2 CLOSED-LOOP PLANNING AND PHASE-II CONTROLS

Overall closed-loop performance. Table 1 compares TrajWAM with reported E2E, VLM, WM, and WAM methods on the Bench2Drive Base Set. Among the listed results, TrajWAM achieves the highest DS (91.53) and SR (75.45%). The next-best results are 89.15 DS from SparseDriveV2 and 72.73% SR from DynVLA, giving margins of 2.38 DS and 2.72 percentage points in SR. Within the

Table 2: Bench2Drive multi-ability results (%). Venue and expert-feature distillation (*) follow Table 1. Bold/underlined values mark the best/second-best listed results.

Method	Venue	Paradigm	Ability (%) ↑					
			Merging	Overtaking	Emergency Brake	Give Way	Traffic Sign	Mean
TCP* (Wu et al., 2022)	NeurIPS '22	E2E	16.18	20.00	20.00	10.00	6.99	14.63
ThinkTwice* (Jia et al., 2023b)	CVPR '23	E2E	27.38	18.42	35.82	**50.00**	54.23	37.17
DriveAdapter* (Jia et al., 2023a)	ICCV '23	E2E	28.82	26.38	48.76	**50.00**	56.43	42.08
AD-MLP (Zhai et al., 2023)	arXiv '23	E2E	0.00	0.00	0.00	0.00	4.35	0.87
UniAD-Base (Hu et al., 2023)	CVPR '23	E2E	14.10	17.78	21.67	10.00	14.21	15.55
VAD (Jiang et al., 2023)	ICCV '23	E2E	8.11	24.44	18.64	20.00	19.15	18.07
DriveTransformer (Jia et al., 2025)	ICLR '25	E2E	17.57	35.00	48.36	40.00	52.10	38.60
SparseDriveV2 (Sun et al., 2026)	ECCV '26	E2E	66.25	75.55	75.00	**50.00**	71.57	67.67
ReasonPlan (Liu et al., 2025)	CoRL '25	VLM	37.50	26.67	33.30	40.00	45.76	36.66
ORION (Fu et al., 2025)	ICCV '25	VLM	25.00	71.11	78.33	30.00	69.15	54.72
SimLingo (Renz et al., 2025)	CVPR '25	VLM	60.00	60.00	78.33	**50.00**	**77.89**	65.25
DeepSight (Zhang et al., 2026)	ICML '26	VLM	60.00	**91.11**	78.33	**50.00**	71.58	**70.20**
UniDrive-WM (Xiong et al., 2026)	ECCV '26	WM	29.97	74.64	79.98	40.00	71.54	59.23
GeRo-Qwen2.5VL (Yasarla et al., 2026)	CVPR '26	WM	37.83	72.81	**83.10**	**50.00**	66.17	61.98
LVDrive (Mei et al., 2026)	arXiv '26	WAM	39.74	69.89	71.67	20.00	74.21	54.90
TrajWAM (Ours)	This work	WAM	**75.00**	82.22	65.00	**50.00**	77.37	69.92

Table 3: Schedule-matched Phase-II comparison on all 220 Bench2Drive routes. Models and objectives follow Section 4.1.

Model	DS↑	SR (%)↑
Base Planner (Phase I)	86.23	71.36%
Base Planner (planning-only FT)	86.22	71.82%
TrajWAM	**91.53**	**75.45%**

Think2Drive WM/WAM group, DynVLA is the strongest prior result on both primary metrics, and TrajWAM improves its DS by 3.19 points and its SR by 2.72 points. The secondary metrics are less uniform: TrajWAM obtains 171.41 Efficiency and 15.27 Smoothness, but leads on neither metric.

Multi-ability performance. Table 2 disaggregates performance across five driving abilities. TrajWAM obtains a mean of 69.92%, ranking second among all listed methods and 0.28 percentage points below the reported DeepSight result. Among WM/WAM entries, the strongest prior mean is 61.98% from GeRo-Qwen2.5VL, giving TrajWAM a descriptive margin of 7.94 percentage points. TrajWAM achieves the highest Merging score (75.00%), ranks second in Overtaking (82.22%) and Traffic Sign (77.37%), and ties the highest Give Way score (50.00%). Its Emergency Brake score is 65.00%, below DeepSight (78.33%) and the best listed result from GeRo-Qwen2.5VL (83.10%), indicating uneven performance across abilities.

Qualitative closed-loop behavior. Figure 3 illustrates how TrajWAM updates its planned trajectory as the scene evolves, alongside the resulting executed trajectory. The low-visibility highway plan shortens as the vehicle brakes for a cut-in, the nighttime sequence shows car-following behavior, and the intersection plan evolves toward a turn as observations update. Planned and executed trajectories can differ because the vehicle continues to replan during execution. These examples illustrate closed-loop adaptation; quantitative comparisons are reported in Tables 3 and 4.

Table 3 compares TrajWAM with Base Planner (planning-only FT) under the same Phase-II training duration. TrajWAM is higher by 5.31 DS points, 5.12 RC points, and 3.64 percentage points in SR (eight additional successful routes). Base Planner (Phase I) is our independently trained and evaluated starting model, distinct from the published DeepSight baseline in Table 1. Table 4 retains the same Base Planner (planning-only FT) control and compares the additional training objectives.

Component ablation. Table 4 compares four models trained independently from Base Planner (Phase I). Base Planner (unconditioned future FT) reaches 73.18% SR (161/220), compared with 71.82% (158/220) for Base Planner (planning-only FT), while DS decreases from 86.22 to 85.34. SR counts strictly successful routes, whereas DS averages route completion after infraction penalties

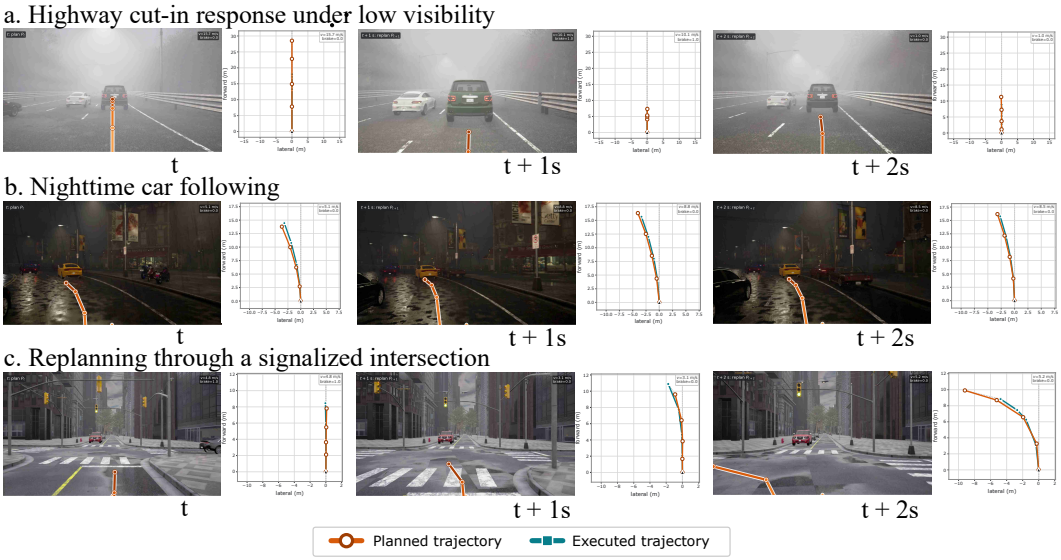

Figure 3: Closed-loop driving examples of TrajWAM on three completed routes selected from the 220-route sweep for scene diversity. At t, $t{+}1\,\mathrm{s}$, and $t{+}2\,\mathrm{s}$, each front-view overlay shows the planned trajectory: four newly predicted waypoints spaced $0.5\,\mathrm{s}$ apart. The adjacent local-BEV plot compares the planned trajectory with the executed trajectory, the vehicle's actual motion over the following $2\,\mathrm{s}$ under continued replanning. Both trajectories are expressed in the ego frame at the planning time.

Table 4: Phase-II component ablation on 220 routes under the setup in Section 4.1. Factual DINO denotes target reconstruction, unconditioned in Base Planner (unconditioned future FT); Traj.-FiLM comprises the ActionEncoder and latent FiLM; Sep. denotes Equation 10. The Base Planner (planning-only FT) control is shared with Table 3.

Model	Factual DINO	Traj.-FiLM	Sep.	DS↑	SR↑
Base Planner (planning-only FT)	–	–	–	86.22	71.82%
Base Planner (unconditioned future FT)	✓	–	–	85.34	73.18%
TrajWAM w/o separation	✓	✓	–	89.34	74.55%
TrajWAM	✓	✓	✓	**91.53**	**75.45%**

over all routes. Three additional strict successes therefore need not yield a higher mean DS: these metrics capture different aspects of performance. Unconditioned future supervision thus improves SR while reducing DS in this comparison.

Adding trajectory conditioning to factual future supervision gives TrajWAM w/o separation 89.34 DS and 74.55% SR (164/220), an increase of 4.00 DS points and three successful routes over Base Planner (unconditioned future FT). Adding separation then gives TrajWAM 91.53 DS and 75.45% SR (166/220), a further increase of 2.19 DS points and two successful routes. These two comparisons report the respective gains of trajectory conditioning and separation in the evaluated training designs. The full model attains the highest DS and SR among the evaluated variants.

Additional diagnostics and footprint. Of 220 routes, 166 are strict successes, 24 reach the destination with a disqualifying infraction, and 30 do not complete; route deviation, collision, and timeout are the most frequent failure signals. Appendix C gives the full outcome accounting, Appendix E.1 defines per-horizon DINO diagnostics, and Appendix D separates checkpoint storage from the native planner-only execution path.

5 LIMITATIONS

Our experiments use a single training seed and one evaluation sweep per variant; robustness across seeds and repeated simulator runs remains to be assessed. Perturbed trajectories have no observed future targets, so the separation objective measures query sensitivity rather than physical counterfactual accuracy. Future-representation visualizations use training examples, and held-out prediction generalization remains untested. The contribution of auxiliary supervision to planning warrants further mechanistic study. Evaluation is limited to CARLA/Bench2Drive; real-world transfer and measured inference latency and FLOPs remain future work.

6 CONCLUSION

TrajWAM learns a trajectory-conditioned World Action Model from single-outcome driving logs by coupling factual DINOv3 BEV reconstruction and unlabeled-query separation with a shared VLM planner. The full objective achieves the highest DS and SR among the evaluated variants in the reported Bench2Drive sweeps, supporting this world–action training design for planning. Native inference retains autoregressive planning and PID control while bypassing the auxiliary readout.

AI USE STATEMENT

No large language models or other AI tools were used in writing this paper. The authors take full responsibility for the content of this work.

ETHICS STATEMENT

This work evaluates autonomous-driving planning exclusively in the CARLA/Bench2Drive simulator and makes no claim of real-world safety. It involves no human subjects or personal data. Deployment on physical vehicles would require independent safety validation, robustness testing, and compliance with applicable regulations.

REPRODUCIBILITY STATEMENT

Section 3 specifies the model and training objectives, Section 4.1 describes the experimental setup, and Appendices A and E provide implementation and evaluation details. The supplementary source-code archive includes the model implementation, runtime patches, and scripts for multi-GPU training, closed-loop evaluation, and metric aggregation. Its README and reproduction guide describe dataset preparation and environment setup, and provide download links for the dataset, trained model weights, and optional recorded closed-loop evaluation outputs.

REFERENCES

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

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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24823–24834. IEEE, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Fu_ORION_A_Holistic_End-to-End_Autonomous_Driving_Framework_by_Vision-Language_Instructed_ICCV_2025_paper.html.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17853–17862. IEEE,

2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Hu_Planning-Oriented_Autonomous_Driving_CVPR_2023_paper.html.

Feiyang Jia, Lin Liu, Ziying Song, Caiyan Jia, Hangjun Ye, Xiaoshuai Hao, and Long Chen. DriveWorld-VLA: Unified latent-space world modeling with vision-language-action for autonomous driving. arXiv preprint arXiv:2602.06521, 2026. URL https://arxiv.org/abs/2602.06521.

Xiaosong Jia, Yulu Gao, Li Chen, Junchi Yan, Patrick Langechuan Liu, and Hongyang Li. Driveadapter: Breaking the coupling barrier of perception and planning in end-to-end autonomous driving. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7919–7929. IEEE, 2023a. URL https://openaccess.thecvf.com/content/ICCV2023/html/Jia_DriveAdapter_Breaking_the_Coupling_Barrier_of_Perception_and_Planning_in_ICCV_2023_paper.html.

Xiaosong Jia, Penghao Wu, Li Chen, Jiangwei Xie, Conghui He, Junchi Yan, and Hongyang Li. Think twice before driving: Towards scalable decoders for end-to-end autonomous driving. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21983–21994. IEEE, 2023b. URL https://openaccess.thecvf.com/content/CVPR2023/html/Jia_Think_Twice_Before_Driving_Towards_Scalable_Decoders_for_End-to-End_Autonomous_CVPR_2023_paper.html.

Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2Drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 819–844. Curran Associates, Inc., 2024. doi: 10.52202/079017-0025. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/017761f94a1cd66d01c041aff85492c4-Abstract-Datasets_and_Benchmarks_Track.html.

Xiaosong Jia, Junqi You, Zhiyuan Zhang, and Junchi Yan. Drivetransformer: Unified transformer for scalable end-to-end autonomous driving. In *International Conference on Learning Representations*, volume 2025, pp. 67227–67243, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a7afc9957f1190223763b6ea93218f98-Abstract-Conference.html.

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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8306–8316. IEEE, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Jiang_VAD_Vectorized_Scene_Representation_for_Efficient_Autonomous_Driving_ICCV_2023_paper.html.

Baoyu Li, Xinchen Yin, Mengying Lin, Yixin Zhang, and Danfei Xu. EgoWAM: World action models beyond pixels with in-the-wild egocentric human data. arXiv preprint arXiv:2607.08436, 2026. URL https://arxiv.org/abs/2607.08436.

Yingyan Li, Lue Fan, Jiawei He, Yuqi Wang, Yuntao Chen, Zhaoxiang Zhang, and Tieniu Tan. Enhancing end-to-end autonomous driving with latent world model. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=fd2u60ryG0.

Yingyan Li, Shuyao Shang, Weisong Liu, Bing Zhan, Haochen Wang, Yuqi Wang, Yuntao Chen, Xiaoman Wang, Yasong An, Chufeng Tang, Lu Hou, Lue Fan, and Zhaoxiang Zhang. DriveVLA-W0: World models amplify data scaling law in autonomous driving. arXiv preprint arXiv:2510.12796, 2025b. URL https://arxiv.org/abs/2510.12796.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27137–27146, October 2025c. URL https://openaccess.thecvf.com/content/ICCV2025/html/

Li_End-to-End_Driving_with_Online_Trajectory_Evaluation_via_BEV_World_Model_ICCV_2025_paper.html.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12037–12047. IEEE, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Liao_DiffusionDrive_Truncated_Diffusion_Model_for_End-to-End_Autonomous_Driving_CVPR_2025_paper.html.

Xueyi Liu, Zuodong Zhong, Yuxin Guo, Yun-Fu Liu, Zhiguo Su, Qichao Zhang, Junli Wang, Yinfeng Gao, Yupeng Zheng, Qiao Lin, et al. Reasonplan: Unified scene prediction and decision reasoning for closed-loop autonomous driving. *arXiv preprint arXiv:2505.20024*, 2025. URL https://proceedings.mlr.press/v305/liu25e.html.

Xiaodong Mei, Diankun Zhang, Hongwei Xie, Guang Chen, Hangjun Ye, and Dan Xu. Lvdrive: Latent visual representation enhanced vision-language-action autonomous driving model. *arXiv preprint arXiv:2605.22089*, 2026. URL https://arxiv.org/abs/2605.22089.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):3942–3951, 2018. doi: 10.1609/aaai.v32i1.11671. URL https://ojs.aaai.org/index.php/AAAI/article/view/11671.

Katrin Renz, Long Chen, Elahe Arani, and Oleg Sinavski. Simlingo: Vision-only closed-loop autonomous driving with language-action alignment. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11993–12003. IEEE, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Renz_SimLingo_Vision-Only_Closed-Loop_Autonomous_Driving_with_Language-Action_Alignment_CVPR_2025_paper.html.

Shuyao Shang, Bing Zhan, Yunfei Yan, Yuqi Wang, Yingyan Li, Yasong An, Xiaoman Wang, Jierui Liu, Lu Hou, Lue Fan, et al. Dynvla: Learning world dynamics for action reasoning in autonomous driving. *arXiv preprint arXiv:2603.11041*, 2026. URL https://arxiv.org/abs/2603.11041.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. arXiv preprint arXiv:2508.10104, 2025. URL https://arxiv.org/abs/2508.10104.

Ziying Song, Caiyan Jia, Lin Liu, Hongyu Pan, Yongchang Zhang, Junming Wang, Xingyu Zhang, Shaoqing Xu, Lei Yang, and Yadan Luo. Don't shake the wheel: Momentum-aware planning in end-to-end autonomous driving. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22432–22441. IEEE, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Song_Dont_Shake_the_Wheel_Momentum-Aware_Planning_in_End-to-End_Autonomous_Driving_CVPR_2025_paper.html.

Wenchao Sun, Xuewu Lin, Keyu Chen, Zixiang Pei, Xiang Li, Yining Shi, and Sifa Zheng. Sparsedrivev2: Scoring is all you need for end-to-end autonomous driving. *arXiv preprint arXiv:2603.29163*, 2026. URL https://arxiv.org/abs/2603.29163.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39724–39733, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Tan_Latent_Chain-of-Thought_World_Modeling_for_End-to-End_Autonomous_Driving_CVPR_2026_paper.html.

Guoqing Wang, Pin Tang, Xiangxuan Ren, Guodongfang Zhao, Bailan Feng, and Chao Ma. Learning vision-language-action world models for autonomous driving. arXiv preprint arXiv:2604.09059, 2026a. URL https://arxiv.org/abs/2604.09059.

Letian Wang, Chuhan Zhang, Rishabh Kabra, Jasper Uijlings, Steven Waslander, Andrew Zisserman, Joao Carreira, Kaiming He, Misha Andriluka, Eduard Gabriel Bazavan, Andrei Zanfir, and Cristian Sminchisescu. Video generation models are general-purpose vision learners. arXiv preprint arXiv:2607.09024, 2026b. URL https://arxiv.org/abs/2607.09024. ECCV 2026 (author-reported).

Linbo Wang, Yupeng Zheng, Qiang Chen, Shiwei Li, Yichen Zhang, Zebin Xing, Qichao Zhang, Xiang Li, Deheng Qian, Pengxuan Yang, Yihang Dong, Ce Hao, Xiaoqing Ye, Junyu Han, Yifeng Pan, and Dongbin Zhao. Latent-WAM: Latent world action modeling for end-to-end autonomous driving. arXiv preprint arXiv:2603.24581v1, 2026c. URL https://arxiv.org/abs/2603.24581v1.

Siyin Wang, Junhao Shi, Zhaoyang Fu, Xinzhe He, Feihong Liu, Chenchen Yang, Yikang Zhou, Zhaoye Fei, Jingjing Gong, Jinlan Fu, Mike Zheng Shou, Xuanjing Huang, Xipeng Qiu, and Yu-Gang Jiang. World Action Models: The next frontier in embodied AI. arXiv preprint arXiv:2605.12090, 2026d. URL https://arxiv.org/abs/2605.12090.

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 (CVPR)*, pp. 14749–14759, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wang_Driving_into_the_Future_Multiview_Visual_Forecasting_and_Planning_with_CVPR_2024_paper.html.

Thomas Wimmer, Prune Truong, Marie-Julie Rakotosaona, Michael Oechsle, Federico Tombari, Bernt Schiele, and Jan Eric Lenssen. Anyup: Universal feature upsampling. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=Y9UAgPehqo.

Penghao Wu, Xiaosong Jia, Li Chen, Junchi Yan, Hongyang Li, and Yu Qiao. Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. *Advances in Neural Information Processing Systems*, 35:6119–6132, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/286a371d8a0a559281f682f8fbf89834-Abstract.html.

Zhexiao Xiong, Xin Ye, Burhan Yaman, Sheng Cheng, Yiren Lu, Jingru Luo, Nathan Jacobs, and Liu Ren. UniDrive-WM: Unified understanding, planning and generation world model for autonomous driving. arXiv preprint arXiv:2601.04453, 2026. URL https://arxiv.org/abs/2601.04453.

Rajeev Yasarla, Deepti Hegde, Shizhong Han, Hsin-Pai Cheng, Yunxiao Shi, Meysam Sadeghigooghari, Shweta Mahajan, Apratim Bhattacharyya, Litian Liu, Risheek Garrepalli, et al. Generative scenario rollouts for end-to-end autonomous driving. *arXiv preprint arXiv:2601.11475*, 2026. URL https://arxiv.org/abs/2601.11475.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi "Jim" Fan, and Joel Jang. World action models are zero-shot policies. arXiv preprint arXiv:2602.15922, 2026. URL https://arxiv.org/abs/2602.15922.

Quanfu Yu, Xian Wu, Hao Xu, and Liulong Ma. HyWorldVLA: A vision-language-action model with hybrid world modeling for autonomous driving. arXiv preprint arXiv:2607.20988, 2026. URL https://arxiv.org/abs/2607.20988.

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. URL https://arxiv.org/abs/2603.16666.

Jiang-Tian Zhai, Ze Feng, Jinhao Du, Yongqiang Mao, Jiang-Jiang Liu, Zichang Tan, Yifu Zhang, Xiaoqing Ye, and Jingdong Wang. Rethinking the open-loop evaluation of end-to-end autonomous driving in nuscenes. *arXiv preprint arXiv:2305.10430*, 2023. URL https://arxiv.org/abs/2305.10430.

Lingjun Zhang, Changjie Wu, Linzhe Shi, Jiangyang Li, Jiaxin Liu, Lei Yang, Hang Zhang, Mu Xu, and Hong Wang. Deepsight: Long-horizon world modeling via latent states prediction for end-to-end autonomous driving. *arXiv preprint arXiv:2605.10564*, 2026. URL https://arxiv.org/abs/2605.10564.

Weiheng Zhao, Haoyi Jiang, Xin Shi, Liu Liu, Fan Huang, Zhizhong Su, Wei Sui, and Xinggang Wang. Faster-WAM: Efficient inference-time future conditioning for robust world action models. arXiv preprint arXiv:2608.04404, 2026a. URL https://arxiv.org/abs/2608.04404.

Zongchuang Zhao, Xin Zhou, Tianyang Xu, Zhengyang Sun, Kaixuan Zhou, Honglin Li, Dingkang Liang, and Xiang Bai. SimWAM: A simple world action model for end-to-end autonomous driving. arXiv preprint arXiv:2608.07468, 2026b. URL https://arxiv.org/abs/2608.07468.

Wenzhao Zheng, Ruiqi Song, Xianda Guo, Chenming Zhang, and Long Chen. Genad: Generative end-to-end autonomous driving. In *European Conference on Computer Vision*, pp. 87–104. Springer, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/8174_ECCV_2024_paper.php.

Yupeng Zheng, Pengxuan Yang, Zebin Xing, Qichao Zhang, Yuhang Zheng, Yinfeng Gao, Pengfei Li, Teng Zhang, Zhongpu Xia, Peng Jia, XianPeng Lang, and Dongbin Zhao. World4Drive: End-to-end autonomous driving via intention-aware physical latent world model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28632–28642, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zheng_World4Drive_End-to-End_Autonomous_Driving_via_Intention-aware_Physical_Latent_World_Model_ICCV_2025_paper.html.

Ruiguo Zhong, Benshan Ma, Xiaolong Chen, Lang Zhang, Mingyue Feng, Yaonong Wang, Pei Liu, and Jun Ma. DA-WAM: Decision-aligned future latents for driving world models. arXiv preprint arXiv:2608.19085, 2026. URL https://arxiv.org/abs/2608.19085.

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. *Advances in Neural Information Processing Systems*, 38:27920–27956, 2026. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/2843fccca5bedd369a4764848b9bd546-Abstract-Conference.html.

A Additional Implementation Details

DINO readout and query construction. Equations 4–6 give the complete DINO-only readout. The ActionEncoder uses $\sigma = 10$; FiLM is zero-initialized around the identity and broadcasts one trajectory code over all valid query tokens and five horizons. Targets are removed from the VLM input and enter only through the frozen DINOv3 teacher. The valid CLS-and-patch-token mask is shared by prediction and target, with register tokens excluded. For A_t^-, the first stored coordinate receives Gaussian noise with standard deviation 2 and the second is independently scaled in $[0.7, 1.3]$, exactly as stated in Equation 3. The reported configuration uses $m_{\text{sem}} = 0.05$ and $\lambda_{\text{sep}} = 0.5$.

Overview visualization. Figure 1 uses AnyUp with logged BEV images as guidance, followed by shared target-fitted PCA and color scaling. This post-processing is used only for display; PCA colors represent feature projections rather than semantic classes or calibrated confidence.

Gradient routing. Let ϕ_{adapt} denote the final four trainable planner layers and θ the ActionEncoder and latent-FiLM parameters. The reported graph satisfies

$$\nabla_{\phi_{\text{adapt}}}\mathcal{L} = \nabla_{\phi_{\text{adapt}}}(\mathcal{L}_{\text{rec}} + 2\mathcal{L}_{\text{fact}} + 0.5\mathcal{L}_{\text{sep}}), \tag{12}$$

$$\nabla_{\theta}\mathcal{L} = \nabla_{\theta}(2\mathcal{L}_{\text{fact}} + 0.5\mathcal{L}_{\text{sep}}), \qquad \nabla_{\theta}\mathcal{L}_{\text{rec}} = 0. \tag{13}$$

The external trajectory is introduced after planner-logit computation, preventing same-pass forward leakage. Apart from the stop-gradient operators explicitly shown in Equations 7 and 10, both the factual and A_t^--conditioned DINO prediction paths remain connected to ϕ_{adapt}. Consequently, $\mathcal{L}_{\text{fact}}$ and $\mathcal{L}_{\text{sep}}$ update the final four trainable planner layers through the DINO branch.

Optimization and parameter accounting. Training uses AdamW with cosine decay, 10% warm-up, BF16, a random seed of 42, and an effective batch size of 64 across eight A100 GPUs. The final four planner layers use learning rate 2×10^{-5} and the task-specific modules use 2×10^{-4}; 330.8M parameters are trainable during adaptation. The stored Base Planner (Phase I) and TrajWAM checkpoints contain 4,063,126,528 and 4,085,644,803 parameters, respectively. This difference includes all auxiliary tensors stored by the training framework and therefore exceeds the size of the modules needed on the native closed-loop path.

B Additional Future-World Prediction Diagnostics

Figures 4 and 5 provide additional native-feature-space diagnostics for longitudinal and lateral trajectory perturbations. Each five-column panel spans t through $t + 2$ s in 0.5 s increments. From top to bottom, the rows show the frozen-DINOv3 target $Y^{\star}_{t+\tau}$, the checkpoint prediction conditioned on the logged four-waypoint trajectory query $A^{\star}_t$, the prediction obtained when the coordinate-perturbed query A^-_t is passed through the same DINOv3-prediction branch, the token-wise factual error $1-\cos(\widehat{Y}_{t+\tau}(A^{\star}_t), Y^{\star}_{t+\tau})$, and the prediction-to-prediction sensitivity $1-\cos(\widehat{Y}_{t+\tau}(A^-_t), \widehat{Y}_{t+\tau}(A^{\star}_t))$. Here $\tau \in \{0, 0.5, 1, 1.5, 2\}$ s. The word "action" in the diagnostic metadata therefore denotes the four-waypoint trajectory query rather than a low-level control command. Because no logged target exists for A^-_t, the final row measures only whether the predicted representation responds to the changed query; it is not counterfactual accuracy.

All displayed future-world maps are outputs of the TrajWAM checkpoint on logged training examples. Within each panel, a single PCA basis and color range are fitted jointly to the five factual DINOv3 targets and then applied unchanged to the targets and both prediction sequences. The underlying tensors remain $16\times16\times1024$; factual error and query sensitivity are computed in the native 1024-dimensional feature space before PCA, and nearest-neighbor enlargement is used only for display. AnyUp is not used in these figures. Consequently, color is comparable across rows and horizons within a panel, but not across independently projected panels.

Examples are selected from logged training data to cover adverse weather, nighttime driving, and both perturbation types. For each frame, the displayed perturbation lies closest to the 75th percentile among 16 fixed seeds rather than the maximum. The set includes the elevated first-horizon factual error in the rainy lateral example. These figures visualize representation alignment and trajectory-query sensitivity; Table 7 reports aggregate factual-target metrics.

C Closed-Loop Failure Analysis

Of 190 destination-reaching routes, 24 fail strict success because of a non-minimum-speed infraction: nine route deviations, eight vehicle collisions, five outside-route-lane violations, one red-light violation, and one layout collision. The other 30 routes terminate as shown in Table 5. Across all 54 non-success routes, the most frequent route-level signals are route deviation (21), vehicle collision (13), and timeout (12); these counts overlap because a route may contain several infractions. The Wilson 95% interval for $166/220$ is 69.36–80.67%, which captures finite-route uncertainty but not training-seed or simulator-repeat variance.

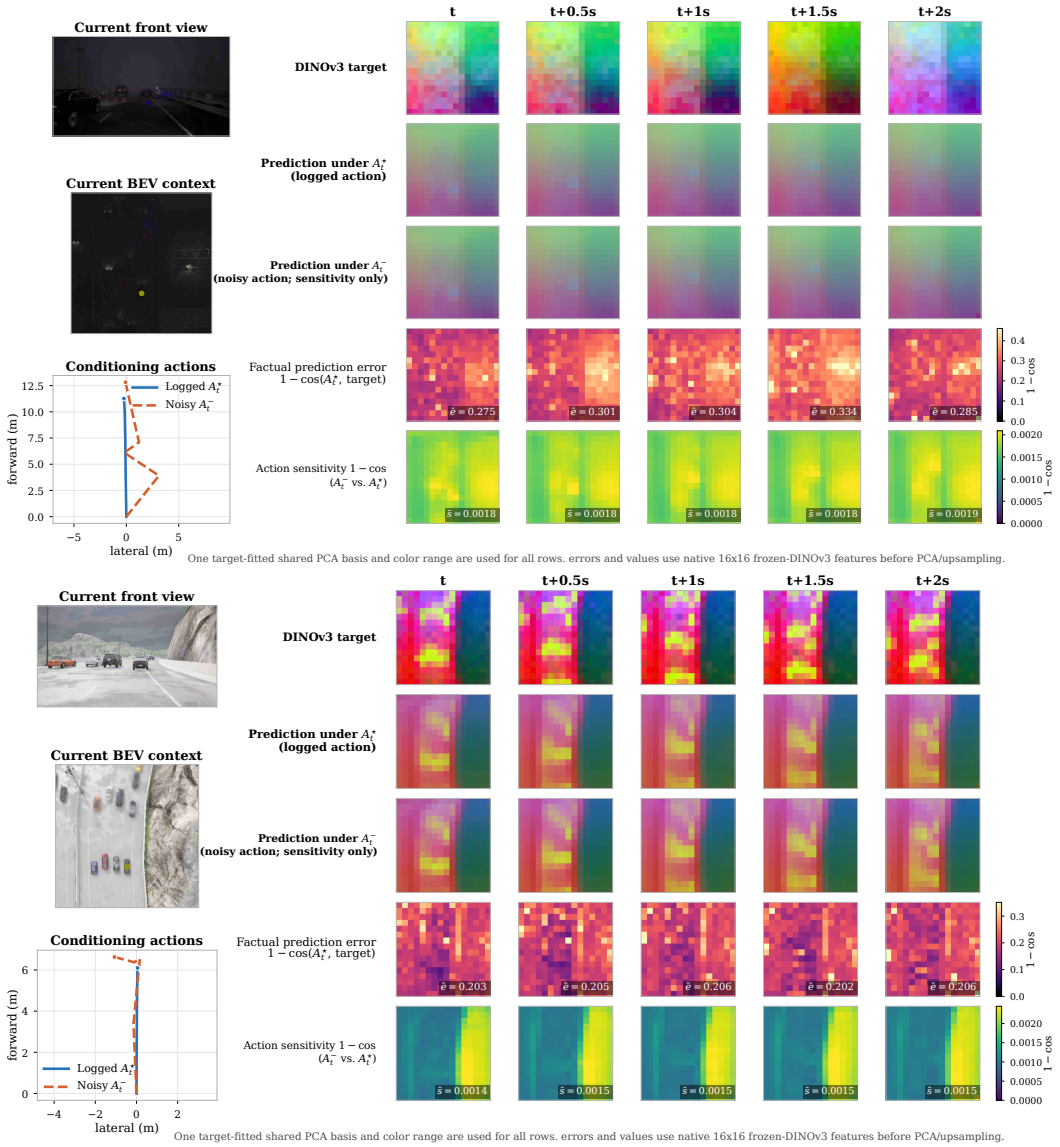

Figure 4: Additional future-world predictions under longitudinal deceleration perturbations. *Top:* rainy example from `Accident_Town04_Route205_Weather23`, frame 51, perturbation seed 2042. *Bottom:* snowy example from `Accident_Town04_Route160_Weather3`, frame 51, seed 2029. In each panel, columns are the five future offsets and rows are the factual DINOv3 target, predictions under $A_t^\star$ and A_t^-, factual cosine error, and label-free query sensitivity, respectively. The maps are outputs of the TrajWAM checkpoint visualized with one target-fitted shared PCA transform per panel; native-space quantities are computed before visualization. A_t^- has no counterfactual target.

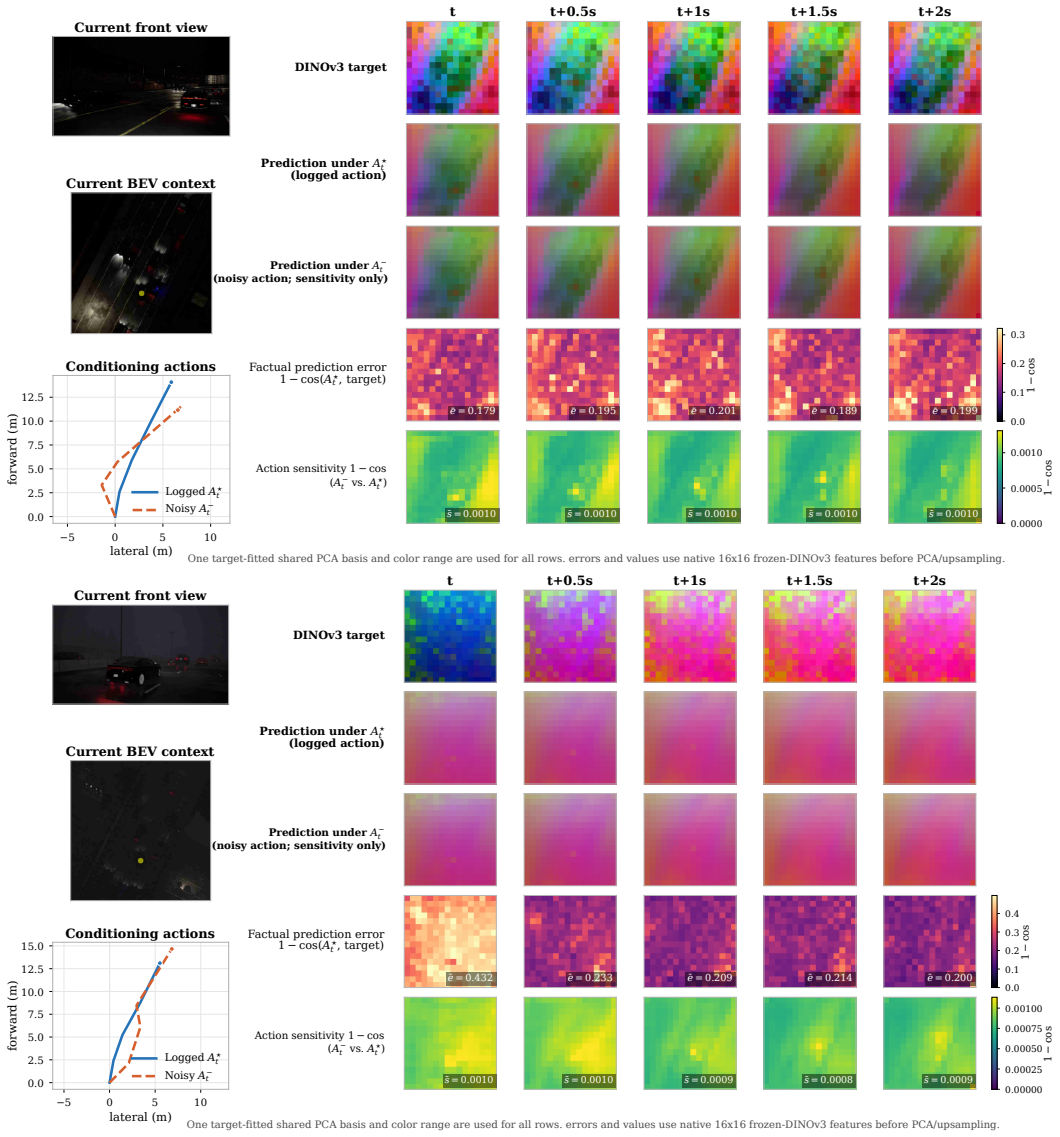

Figure 5: Additional future-world predictions under lateral coordinate perturbations. *Top:* nighttime example from `Accident_Town03_Route102_Weather20`, frame 99, perturbation seed 2033. *Bottom:* rainy example from `Accident_Town04_Route205_Weather23`, frame 163, seed 2040. The five columns and five rows follow Figure 4. All maps are outputs of the TrajWAM checkpoint; each panel uses one target-fitted shared PCA transform and shared display range, while error and sensitivity are computed in native DINOv3 space. Nearest-neighbor enlargement is display-only, AnyUp is not used, and the A_t^- row measures query sensitivity rather than counterfactual accuracy.

Table 5: Mutually exclusive outcomes on all 220 routes. The four termination reasons partition the 30 incomplete routes.

Outcome	Routes	Fraction
Strict success	166	75.45%
Completed with disqualifying infraction	24	10.91%
Did not complete	30	13.64%
Route deviation	12	5.45%
Route timeout	12	5.45%
Vehicle blocked	5	2.27%
Agent crash	1	0.45%

Table 6: Checkpoint footprint and native execution path; decimal units. Training checkpoints include auxiliary tensors that are bypassed online.

Quantity	Base Planner (Phase I)	TrajWAM
Parameters	4.063B	4.086B (+0.554%)
Payload	8.755GB	8.800GB (+45.04MB)
Native path	Planner + PID	Planner + PID

D Checkpoint Storage and Native Execution

The TrajWAM checkpoint contains 22.52M additional parameters. At closed-loop inference, the trajectory-conditioned DINO branch is bypassed, and both models execute autoregressive waypoint decoding and PID control. We report this native-path equivalence without a latency or FLOP comparison.

E Additional Evaluation Details

E.1 DINOv3 Feature-Prediction Protocol

For the factual logged query $A_t^\star$, we evaluate the predicted representation against the frozen DINOv3 target separately at each offset $\tau \in \mathcal{T}$. Let N be the number of diagnostic samples, let $C = 1024$, and let Ω_n contain the valid CLS and spatial patch tokens for sample n after excluding the four register tokens. We compute

$$\begin{aligned} \mathrm{MSE}_\tau &= \frac{1}{C \sum_{n=1}^{N} |\Omega_n|} \sum_{n=1}^{N} \sum_{i \in \Omega_n} \sum_{c=1}^{C} \left(\widehat{Y}_{n,t,\tau,i,c}(A_t^\star) - Y^\star_{n,t,\tau,i,c} \right)^2, \\ \mathrm{Cos}_\tau &= \frac{1}{\sum_{n=1}^{N} |\Omega_n|} \sum_{n=1}^{N} \sum_{i \in \Omega_n} \frac{\widehat{Y}_{n,t,\tau,i}(A_t^\star)^\top Y^\star_{n,t,\tau,i}}{\|\widehat{Y}_{n,t,\tau,i}(A_t^\star)\|_2 \, \|Y^\star_{n,t,\tau,i}\|_2}. \end{aligned} \tag{14}$$

Both metrics are computed on the original 1024-dimensional features, before shared PCA or AnyUp visualization. MSE follows the representation-space training objective, while token-wise cosine similarity separately measures directional alignment. These factual-target metrics do not evaluate the physical correctness of an unobserved alternative future.

Table 7 shows increasing MSE with prediction horizon on the logged future: MSE increases monotonically from 0.030 at t to 0.050 at $t{+}2.0$, while token cosine decreases overall from 0.980 to 0.966, with a small final-offset fluctuation. These results quantify factual-target feature reconstruction only; they do not measure the correctness of unobserved alternative futures.

Route-outcome accounting. Table 5 summarizes all 220 official route records, each with a unique route identifier. We follow the official strict-success predicate: a route must have status `Completed` or `Perfect` and must contain no infraction outside the ignored minimum-speed category. The three

Table 7: Per-offset factual-target prediction in frozen DINOv3 space. MSE uses raw 1024-dimensional channels, and cosine similarity is averaged over valid CLS and patch tokens. The mean is the arithmetic average of the five displayed offsets, rounded to three decimals.

Metric	t	t+0.5	t+1.0	t+1.5	t+2.0	Mean
DINO MSE ↓	0.030	0.033	0.037	0.042	0.050	0.038
Token cosine ↑	0.980	0.980	0.970	0.965	0.966	0.972

top-level outcomes in Table 5 are mutually exclusive; the additional route-level failure-signal counts in the accompanying text may overlap.