MOMWORLD: MOMENTUM-AWARE LATENT WORLD MODEL FOR LONG-HORIZON AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon planning enables autonomous vehicles to anticipate scene evolution and potential risks, supporting safe and stable decisions in complex interactions. However, existing methods struggle to propagate motion trends from observed history into the future. Long rollouts based on a single latent state may further attenuate useful dynamics, retain stale motion patterns, and disrupt reliable near-term plans. We introduce **MomWorld**, a momentum-aware latent world model for long-horizon planning. MomWorld extracts scene motion trends from historical-to-current observations and propagates latent momentum into future horizons, jointly predicting future configuration and momentum states. A learnable momentum persistence mechanism preserves stable trends, scene-conditioned momentum updates adapt future dynamics, and a scene-adaptive reset gate suppresses stale momentum under abrupt changes. We further propose MoFlow, a momentum-conditioned flow-matching module that refines a base trajectory to align with the predicted future scene evolution in only a few integration steps, with a horizon-aware residual fusion that preserves near-term planning stability while permitting stronger long-range corrections. Extensive experiments on NAVSIM, nuScenes and Bench2Drive demonstrate that MomWorld improves long-horizon planning consistency and reduces the average collision rate by 12.2% relative to MomAD over a 6-second planning horizon.

1 INTRODUCTION

Long-horizon planning has become a central frontier in autonomous driving because safe decisions depend not only on the current scene, but also on how traffic may evolve over the next several seconds. Planning-oriented end-to-end systems have advanced this goal by jointly optimizing perception, prediction, and planning within a shared representation (Hu et al., 2023; Jiang et al., 2023; Sun et al., 2025b). More recent history-aware methods extend temporal context, while generative planners enlarge the set of feasible future behaviors (Zhang et al., 2025a; Song et al., 2025; Liao et al., 2025; Song et al., 2026b). Despite this progress, increasing the horizon remains difficult because rollout errors accumulate, interactions such as cut-ins and sudden braking can quickly invalidate earlier assumptions, and uncertain long-range predictions may perturb otherwise reliable near-term decisions. Effective long-horizon planning therefore requires a temporally coherent model of scene evolution rather than a sequence of independently predicted states.

Latent world models offer a promising route to this objective. By forecasting compact, task-relevant future representations, they support planning-oriented imagination without reconstructing every future observation, making multi-step prediction both efficient and directly useful for decision making (Min et al., 2024; Li et al., 2025b; Zheng et al., 2025; Song et al., 2026a). However, existing planning-oriented latent world models generally do not explicitly factorize scene configuration and motion trend into jointly propagated latent states. Repeated transitions may consequently attenuate useful dynamics during smooth motion, retain stale patterns after abrupt interactions, or repeatedly infer the same motion cues from scratch. Moreover, conditioning every planning step on uncertain imagined futures can compromise accurate near-term decisions. **The key challenge is therefore to preserve stable trends, refresh them when the scene changes, and introduce future imagination into planning in a controlled manner.**

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.

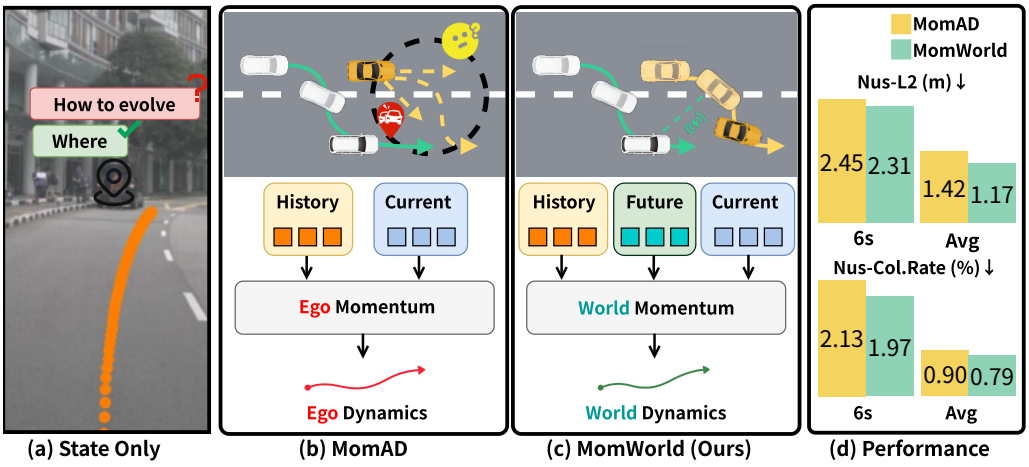


Figure 1: **Motivation and comparison of MomWorld.** (a) Reliable future scene evolution remains a key challenge for long-horizon planning. (b) MomAD (Song et al., 2025) derives ego-centric momentum from historical and current evidence to stabilize ego dynamics. (c) MomWorld elevates momentum to a world-level latent state, initializes it from past-to-present evidence, and jointly rolls out future configuration and momentum to model world dynamics. (d) On six-second nuScenes planning, MomWorld reduces L2 from 2.45 to 2.31 m and collision rate from 2.13% to 1.97%. The corresponding averages improve from 1.42 to 1.17 m and from 0.90% to 0.79%.

MomAD (Song et al., 2025) provides an important insight. Its trajectory and perception momentum connect historical planning and spatiotemporal evidence to the current decision, improving temporal continuity between the past and present. This motivates us to ask whether momentum can connect not only history and the current scene, but also the predicted future. These momentum mechanisms operate within perception and planning queries to stabilize the current ego plan. They do not constitute an explicit latent world state jointly propagated with future scene configurations. As illustrated in Figure 1, extending this idea into future world dynamics can distinguish similar current configurations that imply different evolutions, while allowing obsolete motion trends to be suppressed when an interaction changes abruptly.

We introduce **MomWorld**, a momentum-aware latent world model for long-horizon planning. MomWorld initializes latent scene state and momentum from historical-to-current observations, then jointly propagates them across the predicted future without using future observations at inference. Its first innovation is a scene-adaptive momentum rollout. Learnable persistence retains stable trends, scene-conditioned innovations introduce new dynamics, and a scene-change gate suppresses stale momentum under abrupt interactions. Its second innovation is **MoFlow**, a momentum-conditioned flow-matching module (Lipman et al., 2023) that refines a strong base trajectory toward futures consistent with the predicted scene evolution in only a few integration steps. Bounded horizon-aware residual fusion preserves near-term stability while permitting larger long-range corrections. Experiments on NAVSIM (Dauner et al., 2024; Cao et al., 2025) and nuScenes (Caesar et al., 2020) demonstrate that MomWorld improves long-horizon planning consistency and reduces the average predicted collision rate across the 1–6 s nuScenes horizon by 12.2% relative to MomAD.

We make four contributions.

- We propose **MomWorld**, which connects observed history with future imagination through Latent World Rollout (LWR).
- We introduce **MoLWM**, which maintains, updates, and resets latent momentum to model future dynamics and construct Future World Memory.
- We develop **MoFlow**, which translates predicted world evolution into bounded, horizon-aware trajectory corrections for stable planning.

- We conduct extensive evaluations on NAVSIM, nuScenes, and Bench2Drive demonstrating stronger long-horizon consistency and a 12.2% relative reduction in average predicted collision rate across the 1–6 s nuScenes horizon over MomAD.

2 Related Work

2.1 End-to-End Autonomous Driving

End-to-end autonomous driving jointly learns perception, prediction, and planning around a shared driving objective. UniAD and VAD establish unified planning-oriented architectures with query-based and vectorized scene representations (Hu et al., 2023; Jiang et al., 2023). Recent methods improve computational efficiency through sparse or parallel designs (Sun et al., 2025b; Weng et al., 2024), exploit historical predictions and momentum to enhance temporal consistency (Zhang et al., 2025a; Song et al., 2025), and improve multimodal trajectory generation and selection through truncated diffusion, reinforced diffusion, constraint-guided flow matching, and generalized trajectory scoring (Liao et al., 2025; Song et al., 2026b; Liu et al., 2026a; Li et al., 2025e). These methods primarily improve scene representation, temporal query reuse, or trajectory generation and selection rather than jointly rolling forward an explicit latent state of scene configuration and motion dynamics. MomWorld addresses this gap by rolling past-to-present momentum forward with future latent world states before trajectory refinement.

2.2 Latent World Models for Autonomous Driving

Latent world models encode future scene evolution into compact representations for planning. DriveWorld learns spatiotemporal features with dynamic memory, while LAW predicts ego-trajectory-conditioned future features for self-supervised planning supervision (Min et al., 2024; Li et al., 2025b). World4Drive augments physical latents with spatial-semantic priors and multimodal intentions, whereas Epona jointly models future video and trajectories through autoregressive diffusion (Zheng et al., 2025; Zhang et al., 2025b). Complementary world-model directions focus on human and heterogeneous agent dynamics: Driver-WM causally rolls out in-cabin driver dynamics conditioned on external traffic context (Chi et al., 2026b), whereas PV-WM recurrently co-rolls articulated pedestrians and rigid vehicles within a synchronized structured state (Chi et al., 2026a). Other approaches couple latent imagination with planning more directly: DriveLaW injects video latents into a diffusion planner, DriveFuture conditions diffusion planning on predicted future latents, Latent-WAM autoregressively forecasts compact world states, and GraphWorld transports an ego-centric relational state before modulating motion and planning queries (Xia et al., 2026; Hong et al., 2026; Wang et al., 2026; Song et al., 2026a). In contrast, MomWorld jointly rolls out coupled configuration–momentum states and applies bounded, horizon-aware MoFlow refinement, preserving near-term stability while enabling stronger long-range corrections.

3 Method

3.1 Overview of MomWorld

MomWorld is an end-to-end autonomous-driving framework designed to improve long-horizon planning through momentum-aware latent world modeling. As shown in Fig. 2, historical and current multi-view images are encoded into Scene Queries, with Map/Det heads retained for perception supervision. MomWorld comprises two key components. Momentum-Aware Latent World Modeling uses Latent World Rollout (LWR) to propagate latent scene dynamics and momentum into Future World Memory for candidate scoring and Plan selection. Momentum-Conditioned Flow Matching then uses history-conditioned momentum and predicted future evolution to refine the selected Plan into a temporally consistent trajectory.

3.2 Momentum-Aware Latent World Modeling (MoLWM)

MoLWM represents the latent driving world using a scene state $z_k \in \mathbb{R}^D$ and momentum $p_k \in \mathbb{R}^D$, which encode the scene configuration and its temporal dynamics, respectively. Initialized from his-

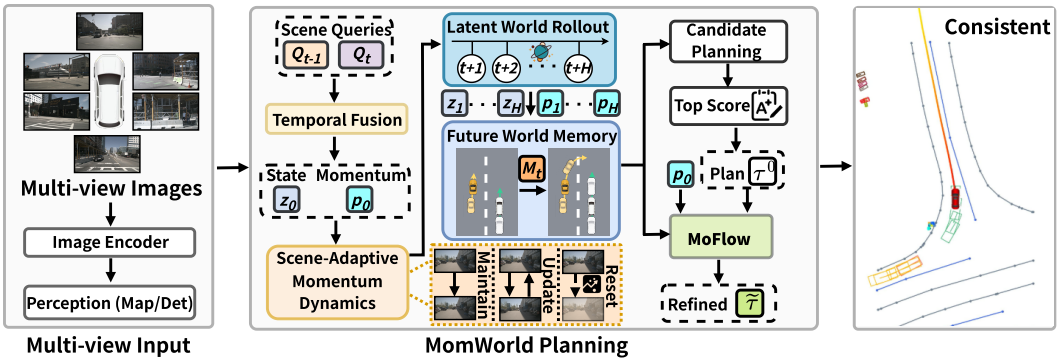


Figure 2: Overview of **MomWorld**. Historical and current multi-view images are encoded into Scene Queries, from which **MoLWM** initializes latent scene state and momentum and propagates them through Latent World Rollout (LWR) to construct Future World Memory for candidate scoring and Plan selection. Guided by history-conditioned momentum and predicted future evolution, **MoFlow** applies bounded residual refinement to produce the temporally consistent Refined trajectory $\widetilde{\tau}$.

torical and current Scene Queries, MoLWM recursively propagates both over the planning horizon to construct Future World Memory for Candidate Planning.

History-Conditioned Initialization. Temporal Fusion combines the current scene representation with its historical variation to initialize z_0 and p_0.

$$(z_0, p_0) = f_{\text{temp}}([\text{Pool}(Q_t), \text{Pool}(Q_t) - \text{Pool}(Q_{t-1})]) \,. \tag{1}$$

The initialized state anchors the current scene configuration, whereas the initialized momentum captures its history-conditioned evolution.

Scene-Adaptive Momentum Dynamics (SMD). To adapt inherited momentum to future scene changes, a multi-head transition network predicts a retention gate ρ_k, a reset gate g_k, and a momentum proposal u_k from the preceding scene state, momentum, and horizon embedding:

$$(\rho_k, g_k, u_k) = f_{\text{gate}}([z_{k-1}, p_{k-1}, r_k^{\text{time}}]). \tag{2}$$

The two gates use sigmoid activations, while the momentum proposal uses a hyperbolic tangent activation. The future momentum is updated by

$$p_k = \text{LN}(\rho_k \odot (1 - g_k) \odot p_{k-1} + (1 - \rho_k) \odot u_k) \,. \tag{3}$$

Here, *Maintain* preserves persistent momentum through ρ_k, *Update* introduces the scene-conditioned proposal u_k, and *Reset* suppresses stale inherited momentum through g_k. Together, these operations produce the scene-adaptive momentum p_k for the k-th future rollout step.

Latent World Rollout (LWR). LWR recursively generates paired future State–Momentum predictions from the history-conditioned pair (z_0, p_0) at the current planning time t. At step k, SMD predicts p_k from (z_{k-1}, p_{k-1}) and the horizon embedding through Eqs. (2)–(3). The predicted momentum then advances the corresponding scene state:

$$z_k = z_{k-1} + \Delta t \, \pi_p(p_k), \qquad k = 1, \ldots, H. \tag{4}$$

Here, π_p maps momentum to a latent transition and Δt is the rollout interval. Feeding each updated pair into the next step produces $\{(z_k, p_k)\}_{k=1}^{H}$. Taking t as the current time index, the k-th pair corresponds to future step $t + k$ and physical offset $k\Delta t$. The state z_k predicts the scene configuration, while p_k encodes history-grounded dynamics adapted to its predicted future evolution.

Algorithm 1 summarizes how training-only future targets ground the rollout. Future observations are encoded into detached target Scene Queries:

$$Q^{\star}_{t+k} = \text{sg}(\text{ImageEncoder}(\mathcal{O}_{t+k})) \,, \qquad k = 1, \ldots, H. \tag{5}$$

Here, $\mathrm{sg}(\cdot)$ blocks gradients through the target branch. The target queries supervise rollout alignment, while future ego, agent, and presence states provide auxiliary supervision. All future targets are used only during training and are unavailable at inference.

Future World Memory (FWM). FWM converts each paired LWR prediction into a memory token aligned with its future step:

$$m_k = \mathrm{LN}\big(z_k + \mathrm{MLP}([p_k, r_k^{\mathrm{time}}])\big), \quad k = 1, \ldots, H. \tag{6}$$

The residual path preserves the predicted scene state z_k, while the learned branch injects its momentum p_k and temporal position r_k^{time}. Each token therefore captures the future scene, dynamics, and horizon information. Dependencies across future steps are modeled by

$$M_t = \mathrm{SelfAttn}([m_1, \ldots, m_H]). \tag{7}$$

Self-attention aggregates context across the rollout, producing a temporally informed memory at each planning step. During Candidate Planning, each candidate query cross-attends to $[Q_t; M_t]$. The scoring heads then rank the decoded candidates and select the Top-Score trajectory as the base Plan τ^0.

Algorithm 1: Future-Supervised Latent World Rollout

Input: Initial state z_0 and momentum p_0; horizon H; future observations $\{\mathcal{O}_{t+k}\}_{k=1}^{H}$; auxiliary targets $\{y^\star_{t+k}\}_{k=1}^{H}$
Output: Future states and momenta $\{(z_k, p_k)\}_{k=1}^{H}$; $\mathcal{L}_{\mathrm{future}}$ and $\mathcal{L}_{\mathrm{aux}}$
Initialize: $\mathcal{L}_{\mathrm{future}}, \mathcal{L}_{\mathrm{aux}} \leftarrow 0$
for $k \leftarrow 1$ **to** H **do**
  Latent World Rollout: obtain (z_k, p_k) using Eqs. (2)–(4)
  Future target: construct $Q^\star_{t+k}$ using Eq. (5)
  Projection: $\widehat{Q}_{t+k} \leftarrow f_Q([z_k, p_k])$
  Alignment: $\mathcal{L}_{\mathrm{future}} \mathrel{+}= d_Q(\widehat{Q}_{t+k}, Q^\star_{t+k})$
  Auxiliary loss: $\mathcal{L}_{\mathrm{aux}} \mathrel{+}= \ell_{\mathrm{aux}}(z_k, p_k; y^\star_{t+k})$
$\mathcal{L}_{\mathrm{future}} \leftarrow \mathcal{L}_{\mathrm{future}}/H, \quad \mathcal{L}_{\mathrm{aux}} \leftarrow \mathcal{L}_{\mathrm{aux}}/H$

f_Q projects each LWR step (z_k, p_k) into query space. d_Q aligns future queries, while ℓ_{aux} supervises future ego, agent, and presence states.

3.3 Momentum-Conditioned Flow Matching (MoFlow)

MoFlow refines the selected Plan τ^0 using the history-conditioned momentum p_0 and Future World Memory M_t. It learns momentum-guided residuals and applies horizon-aware bounded fusion, as shown in Fig. 3, to produce the final trajectory $\widetilde{\tau}$ while preserving near-term stability.

Momentum-Guided Residual Flow (MGRF). The selected Plan τ^0 provides a stable anchor but may not fully capture the predicted historical-to-future dynamics. MoFlow therefore learns a residual vector field conditioned on p_0 and M_t, which provide history-derived motion and horizon-aligned future dynamics, respectively.

The Trajectory Encoder normalizes positions by a fixed scale $s_{\mathrm{pos}} = 50$ and continuously embeds the heading.

$$\phi(x, y, \psi) = (x/s_{\mathrm{pos}}, y/s_{\mathrm{pos}}, \sin\psi, \cos\psi). \tag{8}$$

Applied waypoint-wise, ϕ encodes the base and expert trajectories as $X_0 = \phi(\tau^0)$ and $X_1 = \phi(\tau^\star)$, respectively.

During training, we sample $\epsilon \sim \mathcal{U}(0, 1)$ and construct the linear transport path

$$X_\epsilon = (1 - \epsilon)X_0 + \epsilon X_1. \tag{9}$$

The momentum-conditioned vector field is optimized by

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_\epsilon \left[\| v_\theta(X_\epsilon, \epsilon \mid p_0, M_t) - (X_1 - X_0) \|_2^2 \right]. \tag{10}$$

This formulation learns plan-to-expert residual transport rather than noise-based trajectory generation.

At inference, we initialize the flow state as $X^0 = X_0$ and integrate it using J explicit Euler steps with midpoint flow-time conditioning.

$$X^{j+1} = X^j + \frac{1}{J} v_\theta\left(X^j, \frac{j + 1/2}{J} \mid p_0, M_t\right), \qquad j = 0, \ldots, J - 1. \tag{11}$$

The Trajectory Decoder normalizes the final heading representation and maps it back to trajectory space.

$$\widehat{\tau} = \phi^{-1}\big(\mathcal{N}_\psi(X^J)\big). \tag{12}$$

Horizon-Aware Residual Fusion (HARF). The flow-refined proposal $\widehat{\tau}$ carries historical-to-future corrections, but directly replacing τ^0 may compromise near-term stability. At each horizon step, we first bound its residual.

$$\Delta\tau_k = \mathrm{clip}_{[-\delta,\delta]}\left(\widehat{\tau}_k - \tau_k^0\right). \tag{13}$$

Heading differences are wrapped before clipping. The bounded residual is then fused with the selected Plan.

$$\gamma_k = 0.05 + 0.95\frac{k-1}{H-1}, \qquad \widetilde{\tau}_k = \tau_k^0 + \sigma(b)\gamma_k\Delta\tau_k. \tag{14}$$

Here, δ bounds each residual component, b is a globally learned scalar, and γ_k linearly increases the correction strength from 0.05 to 1.0 toward longer horizons.

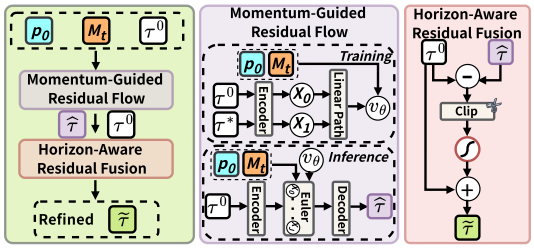


Figure 3: **Momentum-Conditioned Flow Matching (MoFlow).** Conditioned on p_0 and M_t, MoFlow refines the base trajectory through residual flow and bounded horizon-aware fusion.

3.4 Training Objective

Future-world prediction is optimized by

$$\mathcal{L}_{\text{world}} = \mathcal{L}_{\text{future}} + 0.5\mathcal{L}_p + 0.2\mathcal{L}_{\text{aux}}. \tag{15}$$

MoFlow is supervised by

$$\mathcal{L}_{\text{MoFlow}} = \mathcal{L}_{\text{FM}} + \mathcal{L}_\tau. \tag{16}$$

The complete objective is

$$\mathcal{L} = \mathcal{L}_{\text{percep}} + \mathcal{L}_{\text{plan}} + \mathcal{L}_{\text{world}} + \mathcal{L}_{\text{MoFlow}}. \tag{17}$$

Here, $\mathcal{L}_{\text{future}}$, $\mathcal{L}_p$, and $\mathcal{L}_{\text{aux}}$ supervise future queries, momentum, and auxiliary future states, while $\mathcal{L}_{\text{FM}}$ and $\mathcal{L}_\tau$ supervise residual flow and the Refined trajectory. $\mathcal{L}_{\text{percep}}$ and $\mathcal{L}_{\text{plan}}$ are the standard Map/Det and candidate-planning losses. Full definitions are provided in Appendix 1.5.

4 Experiments

4.1 Experimental Setup

Benchmarks. On nuScenes (Caesar et al., 2020), we use the official train/validation split and the camera-only setting. We train separate models for the conventional 3-s horizon (six waypoints) and our 6-s extension (twelve waypoints), both sampled at 2 Hz. The long-horizon model is never obtained by extrapolating a 3-s checkpoint. The 6-s protocol reports planning performance at every second from 1 to 6 s. On NAVSIM v1 (Dauner et al., 2024), we train on *NavTrain* and evaluate all 12,146 indexed scenarios of the official *navtest* split with the classic PDMS. On NAVSIM v2 (Cao et al., 2025), the primary result is the official two-stage *navhard* pseudo-simulation score with corrected human-reference filtering.

Due to space constraints, **Bench2Drive** results are provided in the Appendix.

Metrics. For nuScenes, we report horizon-wise L_2 and predicted collision rate, where lower is better. TPC follows the public MomAD evaluator and measures the dataset-level Euclidean discrepancy between consecutive trajectory predictions (Song et al., 2025). NAVSIM v1 reports PDMS and its components, while NAVSIM v2 reports EPDMS and the navhard stage-wise decomposition, where higher is better.

Implementation Details. For NAVSIM, MomWorld is implemented on GTRS-Dense (Li et al., 2025e) with a V-99-eSE VoVNet backbone and a $D = 256$ latent space. The model takes two camera frames and four LiDAR sweeps, produces 40 LWR steps at $\Delta t = 0.1$ s, and scores a fixed vocabulary of 16,384 trajectories. MoFlow refines the selected trajectory using four explicit Euler steps with midpoint-time conditioning. We train the NAVSIM model for 20 epochs using Adam with a learning rate of 2×10^{-4}, a per-device batch size of 2, mixed precision, and gradient clipping at 5.0. The nuScenes experiments follow the camera-only 3-s and 6-s settings described above. Controlled comparisons use identical inputs, backbones, candidate vocabularies, training data, and optimization budgets within each benchmark.

Table 1: Six-second planning results on the **nuScenes** validation set. Reported FPS uses an A100 for UniAD, an RTX 3090 for LAW, and an RTX 4090 for SparseDrive, MomAD, and GuideFlow.

Method	Venue	L2 (m) ↓							Col. Rate (%) ↓							FPS
		1s	2s	3s	4s	5s	6s	Avg.	1s	2s	3s	4s	5s	6s	Avg.	
UniAD (Hu et al., 2023)	CVPR'23	0.47	0.91	1.35	1.91	2.47	3.07	1.70	0.25	0.36	0.61	0.99	1.64	2.51	1.06	1.8
SparseDrive (Sun et al., 2025b)	ICRA'25	0.43	0.87	1.23	1.75	2.32	2.95	1.59	0.19	0.31	0.56	0.87	1.54	2.33	0.97	9.0
MomAD (Song et al., 2025)	CVPR'25	0.41	0.85	1.13	1.67	1.98	2.45	1.42	0.17	0.30	0.54	0.83	1.43	2.13	0.90	7.8
LAW (Li et al., 2025b)	ICLR'25	0.40	0.87	1.16	1.71	2.03	2.61	1.46	0.19	0.33	0.57	0.86	1.51	2.31	0.96	19.5
Epona (Zhang et al., 2025b)	ICCV'25	0.39	0.91	1.17	1.73	2.02	2.75	1.50	0.14	0.18	0.45	**0.74**	1.48	2.23	0.87	–
World4Drive (Zheng et al., 2025)	ICCV'25	0.42	0.92	1.21	1.75	2.06	2.79	1.53	0.16	0.20	0.47	0.76	1.50	2.14	0.87	–
DIVER (Song et al., 2026b)	TPAMI'26	0.38	0.75	1.10	1.53	1.98	2.49	1.37	0.13	0.31	0.44	0.80	1.41	2.11	0.87	6.6
GuideFlow (Liu et al., 2026a)	CVPR'26	0.42	0.83	1.21	1.73	2.05	2.63	1.48	0.12	0.22	**0.42**	0.79	1.44	2.15	0.86	3.6
MomWorld (Ours)	–	**0.27**	**0.52**	**0.86**	**1.28**	**1.76**	**2.31**	**1.17**	**0.02**	**0.17**	**0.42**	0.83	**1.36**	**1.97**	**0.79**	7.2

4.2 Main Results

Six-second nuScenes planning. Table 1 shows that MomWorld achieves the lowest L2 error at all horizons, reducing the best prior average from 1.37 m to 1.17 m. It also obtains the lowest average collision rate of 0.79%, an 8.1% improvement over the previous best. Overall, MomWorld provides the best long-horizon planning accuracy and safety.

Table 2: Trajectory Prediction Consistency at **4–6 s** on the **nuScenes** validation set.

Method	Venue	TPC (m) ↓			
		4s	5s	6s	Avg.
UniAD (Hu et al., 2023)	CVPR'23	1.49	1.81	2.41	1.90
VAD (Jiang et al., 2023)	ICCV'23	1.55	1.73	2.17	1.82
SparseDrive (Sun et al., 2025b)	ICRA'25	1.33	1.66	1.99	1.66
MomAD (Song et al., 2025)	CVPR'25	1.19	1.45	1.61	1.42
MomWorld (Ours)	–	**0.93**	**1.19**	**1.46**	**1.19**

Long-Horizon Trajectory Prediction Consistency. Table 2 shows that MomWorld achieves the lowest TPC at all horizons, reducing the best prior average from 1.42 m to 1.19 m. Overall, MomWorld provides the best long-horizon trajectory consistency.

NAVSIM v1 navtest. Table 3 shows that MomWorld achieves the best PDMS among world-model methods at 90.2, outperforming DriveLaW by 1.1 points. It also obtains the highest DAC and EP in this category, demonstrating strong planning quality under NAVSIM's non-reactive pseudo-simulation protocol.

NAVSIM v2 navtest. Table 4 compares MomWorld with TransFuser (Chitta et al., 2023), recent end-to-end planners (Liao et al., 2025; Li et al., 2025a; Yao et al., 2025; Zou et al., 2025), VLA methods (Liu et al., 2026b; Li et al., 2025d), and Latent-WAM (Wang et al., 2026). MomWorld achieves the best EPDMS of 90.1, surpassing Latent-WAM by 0.8 points, while attaining the highest TTC and joint-best EP. These results demonstrate that MomWorld achieves a stronger balance between safety, progress, and overall planning quality under the diagnostic pseudo-simulation setting.

NAVSIM v2 navhard. Table 5 compares MomWorld with end-to-end planners (Chitta et al., 2023; Liao et al., 2025; Liu et al., 2026a; Feng et al., 2026; Yao et al., 2025), VLA methods (Dang et al., 2026; Zhou et al., 2026), and world-model methods (Sun et al., 2025a; Zheng et al., 2025). MomWorld achieves the best EPDMS of 42.8 and the highest DDC in both stages, demonstrating stronger robustness under perturbed future observations.

4.3 Ablation Studies

Roles of Different Components in MomWorld. Table 6 shows consistent gains as each component is added. MoLWM progressively improves trajectory accuracy and navhard performance, while MGRF and HARF further reduce long-horizon errors and collisions. The complete MomWorld achieves the best results across all metrics, confirming the complementarity of latent momentum modeling and flow refinement.

Ablation on Different Designs in MoLWM. Table 7 shows that removing any MoLWM design degrades performance, with the momentum proposal and reset gate causing the largest drops. The full MoLWM achieves the best results on all metrics, enabling MomWorld to provide more accurate, consistent, and safe planning.

Ablation on Different Designs in MoFlow. Table 8 studies the conditioning signals, Euler steps, residual fusion designs, and inference latency of MoFlow. We first compare unconditioned refinement with single- and dual-signal conditioning, showing that jointly using p_0 and M_t provides more effective trajectory guidance. We then vary the number of Euler steps, where four steps achieve the

Table 3: Planning performance on the **NAVSIM v1 navtest** split. 'mmt' denotes the multi-modal TransFuser variant, and * marks our re-implementation.

Method	NC↑	DAC↑	TTC↑	Comf.↑	EP↑	PDMS↑
E2E-based Methods						
VADv2 (Chen et al., 2024)	97.2	89.1	91.6	**100**	76.0	80.9
TransFuser (Chitta et al., 2023)	97.7	92.8	92.8	**100**	79.2	84.0
$\text{TransFuser}_{\text{mmt}}$* (Chitta et al., 2023)	96.2	95.4	90.7	**100**	80.7	85.1
UniAD (Hu et al., 2023)	97.8	91.9	92.9	**100**	78.8	83.4
PARA-Drive (Weng et al., 2024)	97.9	92.4	93.0	99.8	79.3	84.0
DRAMA (Yuan et al., 2024)	98.0	93.1	94.8	**100**	80.1	85.5
Hydra-MDP (Li et al., 2024)	98.3	96.0	94.6	**100**	78.7	86.5
FUMP (Liu et al., 2025)	98.1	96.2	94.2	**100**	82.0	87.8
DiffusionDrive (Liao et al., 2025)	98.2	96.2	94.7	**100**	82.2	88.1
DIVER (Song et al., 2026b)	**98.5**	96.5	**94.9**	**100**	82.6	88.3
DriveSuprim (Yao et al., 2025)	97.8	97.3	93.6	**100**	**86.7**	89.9
GoalFlow (Xing et al., 2025)	98.4	**98.3**	94.6	**100**	85.0	**90.3**
ReCogDrive-IL (Li et al., 2025d)	98.1	94.7	94.2	**100**	80.9	86.5
VLA-based Methods						
AutoVLA (Zhou et al., 2025)	98.4	95.6	**98.0**	99.9	81.9	89.1
ReCogdrive (Li et al., 2025d)	98.2	97.8	95.2	99.8	83.5	89.6
DriveWorld-VLA (Liu et al., 2026b)	**99.1**	98.2	96.1	**100**	**85.9**	**91.3**
World-Model-based Methods						
DrivingGPT (Chen et al., 2025)	98.9	90.7	94.9	95.6	79.7	82.4
Resim (Yang et al., 2026a)	–	–	–	–	–	86.6
PWM (Georgiev et al., 2025)	98.6	95.9	95.4	**100**	81.8	88.1
LAW (Li et al., 2025b)	96.4	95.4	88.7	99.9	81.7	84.6
World4Drive (Zheng et al., 2025)	97.4	94.3	92.8	**100**	79.9	85.1
Epona (Zhang et al., 2025b)	97.9	95.1	93.8	99.9	80.4	86.2
WoTE (Li et al., 2025c)	98.5	96.8	94.9	99.9	81.9	88.3
WorldRFT (Yang et al., 2026b)	97.8	96.8	94.0	**100**	81.7	87.8
DriveLaW (Xia et al., 2026)	**99.0**	**97.1**	**96.7**	**100**	81.3	89.1
DriveX-S (Shi et al., 2025)	97.5	94.0	93.0	**100**	79.7	84.5
MomWorld (Ours)	98.6	**98.2**	93.9	**100**	**85.7**	**90.2**

Table 4: Results on the **NAVSIM v2 navtest** split.

Method	NC↑	DAC↑	DDC↑	TLC↑	EP↑	TTC↑	LK↑	HC↑	EC↑	EPDMS↑
E2E-based Methods										
TransFuser	96.9	89.9	97.8	99.7	87.1	95.4	92.7	98.3	87.2	76.7
DiffusionDrive	98.2	95.9	99.4	99.8	87.5	97.3	96.8	98.3	87.7	84.5
Hydra-MDP++	97.2	97.5	99.4	99.6	83.1	96.5	94.4	98.2	70.9	81.4
DriveSuprim	97.5	96.5	99.4	99.6	88.4	96.6	95.5	98.3	77.0	83.1
DiffusionDriveV2	97.7	96.6	99.2	99.8	88.9	97.2	96.0	97.8	91.0	85.5
VLA-based Methods										
DriveWorld-VLA	98.6	99.1	99.6	99.8	87.4	97.9	97.0	97.8	78.6	86.8
ReCogdrive	98.3	95.2	98.3	99.8	87.1	97.5	96.6	99.5	86.5	83.6
World-Model-based Methods										
Latent-WAM	98.1	97.3	99.6	99.8	87.7	97.3	97.6	98.1	87.3	89.3
MomWorld (Ours)	98.1	98.1	99.5	99.8	88.9	98.2	96.1	98.3	90.6	**90.1**

Table 5: Results on the **NAVSIM v2 navhard** split. † computed from reported metrics.

Method	Stage	NC↑	DAC↑	DDC↑	TLC↑	EP↑	TTC↑	LK↑	HC↑	EC↑	EPDMS↑
E2E-based Methods											
TransFuser	Stage 1	96.2	79.5	99.1	99.5	84.1	95.1	94.2	97.5	79.1	23.1
	Stage 2	77.7	70.2	84.2	98.0	85.1	75.6	45.4	95.7	75.9	
DiffusionDrive	Stage 1	96.0	79.7	97.4	99.5	81.3	93.1	90.8	96.8	73.8	24.2
	Stage 2	82.1	72.2	88.5	98.7	85.1	78.8	49.2	89.3	71.2	
GuideFlow	Stage 1	96.6	80.5	96.3	99.3	82.3	94.9	91.5	97.7	67.8	27.1
	Stage 2	87.3	76.7	88.8	99.2	84.3	85.1	49.7	93.1	44.5	
RAP-DINO	Stage 1	97.1	94.4	98.8	99.8	83.9	96.9	94.7	96.4	66.2	36.9
	Stage 2	83.2	83.9	87.4	98.0	86.9	80.4	52.3	95.2	52.4	
DriveSuprim	Stage 1	98.9	95.1	99.2	99.6	76.1	99.1	94.7	97.6	54.2	42.1
	Stage 2	87.9	88.8	89.6	98.8	80.3	86.0	53.5	97.1	56.1	
VLA-based Methods											
DriveFine	Stage 1	97.6	90.0	99.1	99.3	84.9	96.7	97.3	97.6	72.0	30.5†
	Stage 2	82.1	71.3	84.8	98.4	88.1	74.3	47.2	96.8	72.8	
SpanVLA	Stage 1	98.4	94.3	97.8	99.9	85.7	97.2	94.2	97.6	72.1	40.1
	Stage 2	86.9	84.3	87.1	98.2	85.5	82.7	62.3	96.8	67.4	
World-Model-based Methods											
MindDrive	Stage 1	96.1	86.0	98.8	99.3	83.3	95.6	94.4	97.6	74.7	30.9
	Stage 2	82.6	79.1	86.4	98.0	85.3	79.4	49.2	96.5	71.0	
World4Drive	Stage 1	97.3	89.1	97.6	99.7	60.5	96.8	87.7	93.1	60.0	34.9
	Stage 2	91.4	82.0	91.0	98.5	53.1	90.6	52.3	93.3	62.8	
MomWorld (Ours)	Stage 1	96.9	93.6	99.8	99.8	80.4	96.9	96.4	97.6	60.0	**42.8**
	Stage 2	86.4	88.2	94.2	98.3	81.7	84.4	55.3	97.0	54.7	

Table 6: **Roles of Different Method Components in MomWorld.** Cumulative component study on the **nuScenes** validation set and the **NAVSIM v2 navhard** split. FWM, LWR, SMD, MGRF, and HARF denote *Future World Memory*, *Latent World Rollout*, *Scene-Adaptive Momentum Dynamics*, *Momentum-Guided Residual Flow*, and *Horizon-Aware Residual Fusion*, respectively. LK, EP, and EC are the Stage-2 metrics reported by the navhard protocol.

MoLWM			MoFlow		nuScenes 6s					NAVSIM v2 navhard			
FWM	LWR	SMD	MGRF	HARF	L_2@3 (m) ↓	L_2@6 (m) ↓	TPC@6 (m) ↓	Col. @3 (%) ↓	Col. @6 (%) ↓	LK ↑	EP ↑	EC ↑	EPDMS ↑
					1.13	2.45	1.61	0.54	2.13	54.6	69.5	49.7	41.7
✓					1.07	2.42	1.57	0.51	2.09	54.8	72.3	50.8	42.0
✓	✓				1.01	2.39	1.54	0.48	2.06	55.0	75.6	52.1	42.2
✓	✓	✓			0.96	2.36	1.51	0.46	2.03	55.1	78.4	53.0	42.4
✓	✓	✓	✓		0.91	2.33	1.48	0.44	2.00	55.2	80.2	54.0	42.6
✓	✓	✓	✓	✓	**0.86**	**2.31**	**1.46**	**0.42**	**1.97**	**55.3**	**81.7**	**54.7**	**42.8**

best accuracy–efficiency trade-off. Finally, removing residual clipping, horizon-aware weighting, or the learned global gate consistently degrades performance, while direct trajectory replacement performs worst. Overall, the full MoFlow achieves the best results with only 8.4 ms additional latency, enabling MomWorld to deliver more accurate, safe, and efficient planning.

Table 7: Ablation study of different designs in **MoLWM** across **nuScenes**, **NAVSIM v1 navtest**, and **NAVSIM v2 navhard**.

Variant	nuScenes 6s					NAVSIM v1 navtest	NAVSIM v2 navhard
	L_2@3 (m) ↓	L_2@6 (m) ↓	TPC@6 (m) ↓	Col. @3 (%) ↓	Col. @6 (%) ↓	PDMS ↑	EPDMS ↑
No historical variation	0.95	2.58	1.72	0.55	2.40	88.6	39.7
No horizon embedding in SMD	0.90	2.43	1.57	0.47	2.12	89.5	41.3
Fixed retention ($\rho_k \equiv 0.9$)	0.92	2.49	1.63	0.50	2.21	89.0	40.8
No momentum proposal ($u_k \equiv 0$)	0.99	2.69	1.82	0.60	2.58	87.8	38.6
No reset gate ($g_k \equiv 0$)	0.94	2.55	1.69	0.65	2.83	88.2	37.9
No horizon embedding in FWM	0.91	2.46	1.60	0.49	2.17	89.3	40.9
No momentum injection in FWM	0.97	2.64	1.78	0.58	2.49	88.0	38.9
No temporal aggregation in FWM	0.93	2.52	1.66	0.52	2.29	88.8	40.2
Full MoLWM	**0.86**	**2.31**	**1.46**	**0.42**	**1.97**	**90.2**	**42.8**

4.4 VISUALIZATION

Figure 4 presents consecutive planning results for MomAD (Song et al., 2025) and MomWorld. Across $t-1$, t, and $t+1$, MomWorld exhibits smoother trajectory evolution and closer alignment with the ground truth, demonstrating improved temporal stability and planning accuracy.

Table 8: Ablation study of different designs in **MoFlow** across **nuScenes**, **NAVSIM v1 navtest**, and **NAVSIM v2 navhard**.

MoFlow Design	Steps	nuScenes 6s					NAVSIM v1 navtest	NAVSIM v2 navhard	Latency (ms) ↓	
		L_2@3 (m) ↓	L_2@6 (m) ↓	TPC@6 (m) ↓	Col. @3 (%) ↓	Col. @6 (%) ↓	PDMS ↑	EPDMS ↑	Flow	Total
No MoFlow (base plan τ^0)	0	0.96	2.36	1.51	0.46	2.03	89.7	42.4	0.0	130.5
Unconditioned MGRF	4	1.01	2.53	1.67	0.57	2.38	88.3	40.0	8.4	138.9
M_t only	4	0.91	2.40	1.55	0.47	2.09	89.4	41.8	8.4	138.9
p_0 only	4	0.93	2.43	1.58	0.49	2.14	89.2	41.5	8.4	138.9
p_0, M_t	1	0.91	2.38	1.53	0.46	2.07	89.8	42.3	2.2	132.7
p_0, M_t	2	0.88	2.34	1.49	0.44	2.01	90.0	42.5	4.3	134.8
p_0, M_t	8	0.87	2.32	1.47	0.43	1.99	90.1	42.7	16.5	147.0
Direct replacement ($\tilde{\tau} = \hat{\tau}$)	4	0.98	2.57	1.69	0.59	2.46	87.9	39.5	8.4	138.9
No residual clipping	4	0.92	2.45	1.59	0.55	2.37	88.7	40.5	8.4	138.9
No horizon factor	4	0.90	2.40	1.55	0.48	2.12	89.4	41.7	8.4	138.9
No learned global gate ($\sigma(b) \to 1$)	4	0.89	2.38	1.53	0.47	2.09	89.6	42.0	8.4	138.9
Full MoFlow	**4**	**0.86**	**2.31**	**1.46**	**0.42**	**1.97**	**90.2**	**42.8**	8.4	138.9

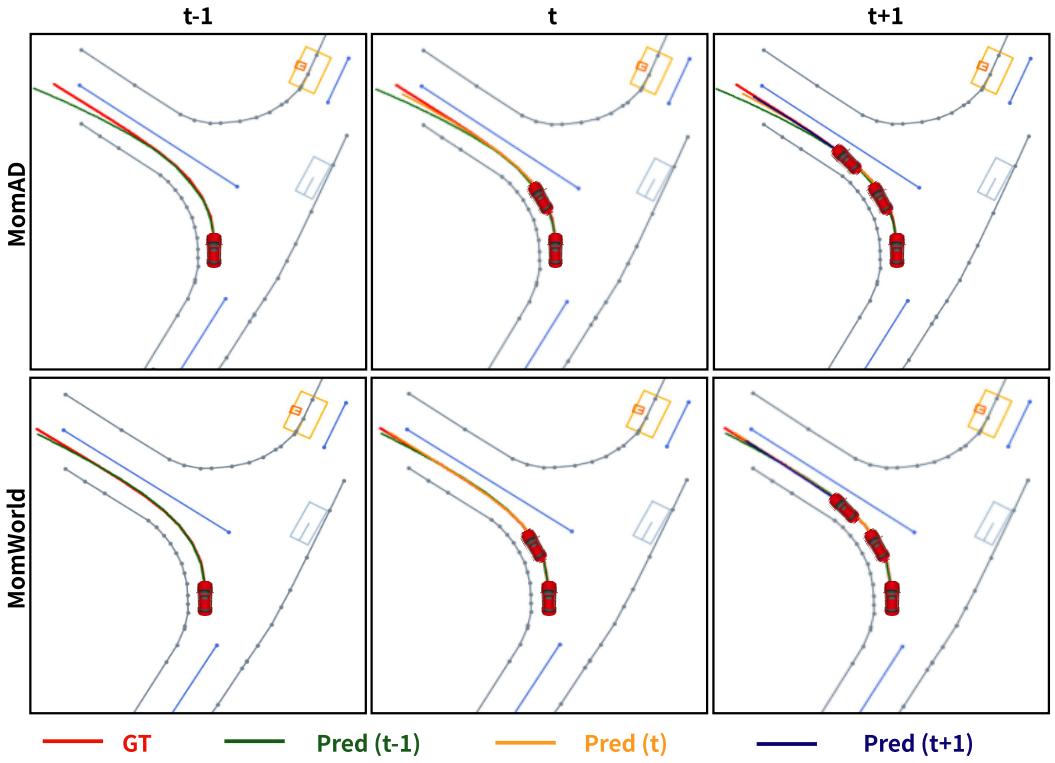


Figure 4: **Consecutive planning visualization on nuScenes.** Comparison of MomAD (Song et al., 2025) and MomWorld across consecutive frames. MomWorld yields more accurate and temporally consistent future trajectories.

5 Conclusion

We presented **MomWorld**, a momentum-aware latent world model for temporally consistent long-horizon autonomous driving. MoLWM jointly propagates future latent scene states and momentum, using scene-adaptive retention, update, and reset mechanisms to preserve persistent dynamics while suppressing stale motion. The resulting Future World Memory supports candidate selection, while MoFlow refines the selected plan through momentum-conditioned residual flow and horizon-aware bounded fusion. Experiments on nuScenes, NAVSIM v1/v2, and Bench2Drive demonstrate improved planning accuracy, temporal consistency, safety, and planning quality across open-loop and pseudo-simulation protocols, confirming the complementarity of latent momentum modeling and controlled flow refinement.

Limitation and Future Work. One limitation of MomWorld is its reliance on a fixed candidate vocabulary and scoring stage. Future work will explore more flexible proposal generation while retaining stable momentum-conditioned refinement.

AI USE STATEMENT

In this work, we used generative AI tools to assist with formulating mathematical claims, providing feedback on experiments, implementing methods, and interpreting results. We did not use generative AI tools to develop theoretical models or conceptual frameworks or to propose or refine hypotheses. The following uses were not applicable to this work: generating synthetic datasets, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, assisting with translation, cleaning or reformatting datasets, and supporting qualitative or thematic data analysis.

Additionally, we used generative AI tools to edit the research paper for improved readability, identify relevant literature, and propose a title for the paper. We reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by the authors, and LLM-generated mathematical claims were examined and refined to ensure that they were correct and aligned with the actual method. 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

Section 3 specifies the MoLWM and MoFlow architecture, state updates, learning objectives, and inference procedure. Section 4.1 describes the datasets, benchmark splits, metrics, and comparison settings. The Appendix provides the model configuration, optimization settings, evaluator definitions, and aggregation rules.

REFERENCES

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11621–11631, 2020.

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 *Proceedings of the 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 4709–4722. PMLR, 2025. URL https://proceedings.mlr.press/v305/cao25a.html.

Shaoyu Chen, Bo Jiang, Hao Gao, Bencheng Liao, Qing Xu, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. VADv2: End-to-end vectorized autonomous driving via probabilistic planning. *arXiv preprint arXiv:2402.13243*, 2024.

Yuntao Chen, Yuqi Wang, and Zhaoxiang Zhang. DrivingGPT: Unifying driving world modeling and planning with multi-modal autoregressive transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 26890–26900, 2025.

Haozhuang Chi, Jingsong Liang, Ziying Song, Lei Yang, Shihao Li, Haoruo Zhang, and Chen Lv. Pv-wm: A heterogeneous micro-macro world model for articulated pedestrian-vehicle co-rollout, 2026a. URL https://arxiv.org/abs/2609.07328.

Haozhuang Chi, Daosheng Qiu, Hao Su, Haochen Liu, Zirui Li, Haoruo Zhang, and Chen Lv. Driver-wm: A driver-centric traffic-conditioned latent world model for in-cabin dynamics rollout, 2026b. URL https://arxiv.org/abs/2605.05092.

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*, 45(11):12878–12895, 2023. doi: 10.1109/TPAMI.2022.3200245.

Chenxu Dang, Sining Ang, Yongkang Li, Haochen Tian, Jie Wang, Guang Li, Hangjun Ye, Jie Ma, Long Chen, and Yan Wang. DriveFine: Refining-augmented masked diffusion vla for precise and robust driving. *arXiv preprint arXiv:2602.14577*, 2026.

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. NAVSIM: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In *Advances in Neural Information Processing Systems*, volume 37, pp. 28706–28719. Curran Associates, Inc., 2024. doi: 10.52202/079017-0902.

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 *International Conference on Learning Representations*, volume 2026, pp. 37722–37737, 2026.

Ignat Georgiev, Varun Giridhar, Nick Hansen, and Animesh Garg. Pwm: Policy learning with multi-task world models. In *International Conference on Learning Representations*, volume 2025, pp. 13737–13757, 2025.

Yufeng Hong, Xiaotian Zhou, Yingyan Li, Xiangpo Zhou, Lin Liu, Yadan Luo, Shaoqing Xu, Lei Yang, and Ziying Song. DriveFuture: Future-aware latent world models for autonomous driving, 2026. URL https://arxiv.org/abs/2605.09701.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17853–17862, 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8350, 2023.

Kailin Li, Zhenxin Li, Shiyi Lan, Yuan Xie, Zhizhong Zhang, Jiayi Liu, Zuxuan Wu, Zhiding Yu, and Jose M Alvarez. Hydra-MDP++: Advancing end-to-end driving via expert-guided hydra-distillation. *arXiv preprint arXiv:2503.12820*, 2025a.

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 *International Conference on Learning Representations*, volume 2025, pp. 42942–42959, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/6aa4967920e495e90aeeaa3acf18d019-Paper-Conference.pdf.

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. *arXiv preprint arXiv:2504.01941*, 2025c.

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. *arXiv preprint arXiv:2506.08052*, 2025d.

Zhenxin Li, Kailin Li, Shihao Wang, Shiyi Lan, Zhiding Yu, Yishen Ji, Zhiqi Li, Ziyue Zhu, Jan Kautz, Zuxuan Wu, et al. Hydra-MDP: End-to-end multimodal planning with multi-target hydra-distillation. *arXiv preprint arXiv:2406.06978*, 2024.

Zhenxin Li, Wenhao Yao, Zi Wang, Xinglong Sun, Joshua Chen, Nadine Chang, Maying Shen, Zuxuan Wu, Shiyi Lan, and Jose M. Alvarez. Generalized trajectory scoring for end-to-end multimodal planning. *arXiv preprint arXiv:2506.06664*, 2025e.

Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, and Xinggang Wang. DiffusionDrive: Truncated diffusion model for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12037–12047, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Lin Liu, Caiyan Jia, Ziying Song, Hongyu Pan, Bencheng Liao, Wenchao Sun, Yongchang Zhang, Lei Yang, and Yandan Luo. Fully unified motion planning for end-to-end autonomous driving. *arXiv preprint arXiv:2504.12667*, 2025.

Lin Liu, Caiyan Jia, Guanyi Yu, Ziying Song, Junqiao Li, Feiyang Jia, Peiliang Wu, Xiaoshuai Hao, and Yadan Luo. GuideFlow: Constraint-guided flow matching for planning in end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3719–3728, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Liu_GuideFlow_Constraint-Guided_Flow_Matching_for_Planning_in_End-to-End_Autonomous_Driving_CVPR_2026_paper.html.

Lin Liu, Ziying Song, Caiyan Jia, Hangjun Ye, Xiaoshuai Hao, Long Chen, et al. DriveWorld-VLA: Unified latent-space world modeling with vision-language-action for autonomous driving. *arXiv preprint arXiv:2602.06521*, 2026b.

Chen Min, Dawei Zhao, Liang Xiao, Jian Zhao, Xinli Xu, Zheng Zhu, Lei Jin, Jianshu Li, Yulan Guo, Junliang Xing, Liping Jing, Yiming Nie, and Bin Dai. DriveWorld: 4d pre-trained scene understanding via world models for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15522–15533, 2024.

Chen Shi, Shaoshuai Shi, Kehua Sheng, Bo Zhang, and Li Jiang. DriveX: Omni scene modeling for learning generalizable world knowledge in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 28599–28609, 2025.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22432–22441, 2025.

Ziying Song, Caiyan Jia, Lin Liu, Lei Yang, Shengkai Zhang, Feiyang Jia, Fengda Zhao, Peiliang Wu, Shaoqing Xu, Chen Lv, and Yadan Luo. GraphWorld: Long-horizon planning with world models for end-to-end autonomous driving, 2026a. URL https://arxiv.org/abs/2606.16274.

Ziying Song, Lin Liu, Hongyu Pan, Bencheng Liao, Mingzhe Guo, Lei Yang, Yongchang Zhang, Shaoqing Xu, Caiyan Jia, and Yadan Luo. DIVER: Reinforced diffusion breaks imitation bottlenecks in end-to-end autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–17, 2026b. doi: 10.1109/TPAMI.2026.3708096.

Bin Sun, Yaoguang Cao, Yan Wang, Rui Wang, Jiachen Shang, Xiejie Feng, Jiayi Lu, Jia Shi, Shichun Yang, Xiaoyu Yan, et al. MindDrive: An all-in-one framework bridging world models and vision-language model for end-to-end autonomous driving. *arXiv preprint arXiv:2512.04441*, 2025a.

Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. SparseDrive: End-to-end autonomous driving via sparse scene representation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8795–8801. IEEE, 2025b. doi: 10.1109/ICRA55743.2025.11128800.

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, 2026. URL https://arxiv.org/abs/2603.24581.

Xinshuo Weng, Boris Ivanovic, Yan Wang, Yue Wang, and Marco Pavone. PARA-Drive: Parallelized architecture for real-time autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15449–15458, 2024.

Tianze Xia, Yongkang Li, Lijun Zhou, Jingfeng Yao, Kaixin Xiong, Haiyang Sun, Bing Wang, Kun Ma, Guang Chen, Hangjun Ye, Wenyu Liu, and Xinggang Wang. DriveLaW: Unifying planning and video generation in a latent driving world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39701–39712, 2026.

Zebin Xing, Xingyu Zhang, Yang Hu, Bo Jiang, Tong He, Qian Zhang, Xiaoxiao Long, and Wei Yin. GoalFlow: Goal-driven flow matching for multimodal trajectories generation in end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1602–1611, 2025.

Jiazhi Yang, Kashyap Chitta, Shenyuan Gao, Long Chen, Yuqian Shao, Xiaosong Jia, Hongyang Li, Andreas Geiger, Xiangyu Yue, and Li Chen. Resim: Reliable world simulation for autonomous driving. *Advances in Neural Information Processing Systems*, 38:167710–167741, 2026a.

Pengxuan Yang, Ben Lu, Zhongpu Xia, Chao Han, Yinfeng Gao, Teng Zhang, Kun Zhan, XianPeng Lang, Yupeng Zheng, and Qichao Zhang. WorldRFT: Latent world model planning with reinforcement fine-tuning for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11649–11657, 2026b.

Wenhao Yao, Zhenxin Li, Shiyi Lan, Zi Wang, Xinglong Sun, Jose M Alvarez, and Zuxuan Wu. DriveSuprim: Towards precise trajectory selection for end-to-end planning. *arXiv preprint arXiv:2506.06659*, 2025.

Chengran Yuan, Zhanqi Zhang, Jiawei Sun, Shuo Sun, Zefan Huang, Christina Dao Wen Lee, Dongen Li, Yuhang Han, Anthony Wong, Keng Peng Tee, et al. DRAMA: An efficient end-to-end motion planner for autonomous driving with mamba. *arXiv preprint arXiv:2408.03601*, 2024.

Bozhou Zhang, Nan Song, Xin Jin, and Li Zhang. Bridging past and future: End-to-end autonomous driving with historical prediction and planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6854–6863, 2025a.

Kaiwen Zhang, Zhenyu Tang, Xiaotao Hu, Xingang Pan, Xiaoyang Guo, Yuan Liu, Jingwei Huang, Li Yuan, Qian Zhang, Xiao-Xiao Long, Xun Cao, and Wei Yin. Epona: Autoregressive diffusion world model for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27220–27230, 2025b. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_Epona_Autoregressive_Diffusion_World_Model_for_Autonomous_Driving_ICCV_2025_paper.html.

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*, pp. 28632–28642, 2025.

Zewei Zhou, Tianhui Cai, Seth Z. 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. *arXiv preprint arXiv:2506.13757*, 2025.

Zewei Zhou, Ruining Yang, Yiluan Guo, Sherry X Chen, Tao Feng, Kateryna Pistunova, Yishan Shen, Lili Su, Jiaqi Ma, et al. SpanVLA: Efficient action bridging and learning from negative-recovery samples for vision-language-action model. *arXiv preprint arXiv:2604.19710*, 2026.

Jialv Zou, Shaoyu Chen, Bencheng Liao, Zhiyu Zheng, Yuehao Song, Lefei Zhang, Qian Zhang, Wenyu Liu, and Xinggang Wang. DiffusionDriveV2: Reinforcement learning-constrained truncated diffusion modeling in end-to-end autonomous driving. *arXiv preprint arXiv:2512.07745*, 2025.