MINDDRIVE: AN ALL-IN-ONE FRAMEWORK BRIDGING WORLD MODEL AND VISION-LANGUAGE MODEL FOR END-TO-END AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

End-to-end autonomous driving (E2E-AD) has emerged as a new paradigm, where trajectory planning plays a crucial role. Existing studies mainly follow two directions: **trajectory generation-oriented** approaches, which focus on producing high-quality trajectories with simple decision mechanisms, and **trajectory selection-oriented** approaches, which perform multi-dimensional evaluation to select the best trajectory yet lack sufficient generative capability. In this work, we propose MindDrive, a harmonized framework that integrates high-quality trajectory generation with comprehensive decision reasoning. It establishes a structured reasoning paradigm of "what-if simulation – candidate generation – multi-objective trade-off." In particular, the proposed **Future-aware Trajectory Generator (FaTG)**, based on a **World Action Model** (WAM), performs ego-conditioned "what-if" simulations to predict potential future scenes and generate foresighted trajectory candidates. Building upon this, the **VLM-oriented Evaluator (VLoE)** leverages the reasoning capability of a large **vision–language model** to conduct multi-objective evaluations across safety, comfort, and efficiency dimensions, leading to reasoned and human-aligned decision-making. Extensive experiments on the NAVSIM-v1 and NAVSIM-v2 benchmarks demonstrate that MindDrive achieves state-of-the-art performance across multi-dimensional driving metrics, significantly enhancing safety, compliance, and generalization, and offering a promising path toward interpretable and cognitively guided autonomous driving.

1 INTRODUCTION

End-to-end autonomous driving (E2E-AD) has made remarkable progress by unifying perception, prediction, and planning within a differentiable framework, reducing error propagation and improving cross-task coordination. Early research focused on perception (Chen & Krähenbühl, 2022; Song et al., 2023; Chitta et al., 2023; Yang et al., 2024), while Think Twice (Jia et al., 2023) highlighted the importance of the planning decoder's architecture, capacity, and reasoning ability. UniAD (Hu et al., 2023) integrates multiple sub-tasks into a single training pipeline through a dense perception–prediction–planning cascade. Subsequent lightweight methods use vectorized or sparse representations to model lanes, intersections, and dynamic agents at lower computational cost. Recent studies combine generative paradigms with Transformer-based planning heads sharing spatio-temporal features to improve trajectory quality. As perception modules mature, research has increasingly shifted toward efficient, safe, and comfortable trajectory planning. Existing planning studies fall into two main categories: trajectory generation and trajectory selection.

Trajectory generation-oriented approaches (Fig. 1(a)) improve trajectory diversity and adaptability. VADv2 (Chen et al., 2024a), GaussianAD (Zheng et al., 2024b), and DiffusionDrive (Liao et al., 2025) use multimodal autoregressive modeling, probabilistic distributions, or diffusion to capture driving uncertainty and complex traffic dynamics. Bridging Past and Future (Zhang et al., 2025) and MomAD (Song et al., 2025a) exploit historical motion cues to improve temporal coherence and consistency with long-term driving intent. Large language and vision–language models further incorporate high-level semantics and scene reasoning into trajectory generation (Jiang et al., 2024; Xing et al., 2025a; Zhou et al., 2025).

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.

Trajectory selection-oriented approaches (Fig. 1(b)) evaluate pre-generated candidates to select the most feasible trajectory. WoTE (Li et al., 2025b) and Hydra-MDP (Li et al., 2024c) train neural scorers to approximate dataset-defined benchmark scores, replacing hand-crafted evaluators. SimpleVSF (Zheng et al., 2025a) further introduces VLM-enhanced scoring for richer contextual and semantic understanding.

Despite these advances, **trajectory generation-oriented methods** typically select candidates using simple MLP heads, heuristics, or coarse scoring strategies, lacking comprehensive evaluation of safety, comfort, efficiency, and rule compliance. They may therefore fail to select an excellent trajectory even when one is available. Conversely, **trajectory selection-oriented methods** pair sophisticated scorers with traditional or lightweight generators, leaving performance constrained by candidate coverage and quality—**a strong evaluator cannot rescue a weak generator**. Although both follow a generate–then–select pipeline, their emphasis on one stage creates **an inherent imbalance between the two stages—a core limitation of existing planning pipelines**. This imbalance limits unified modeling and joint decision making, motivating a more balanced and integrated planning framework.

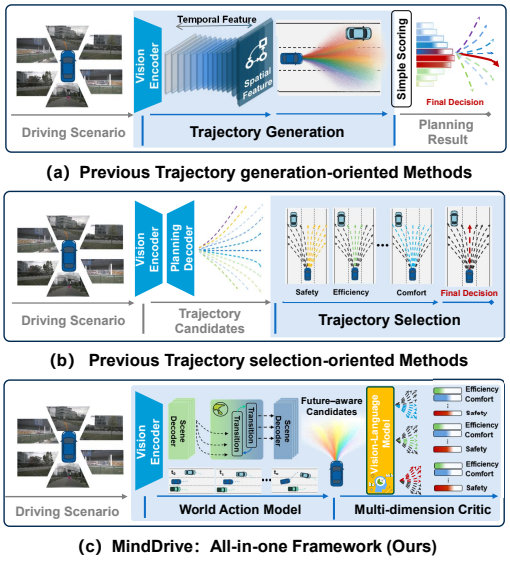


Figure 1: **Comparison of planning paradigms.** **(a)** Generation-oriented methods (Hu et al., 2023; Chen et al., 2024a; Sun et al., 2025; Song et al., 2025a) generate diverse trajectories but rely on weak selectors. **(b)** Selection-oriented methods (Li et al., 2024c; 2025b; Zheng et al., 2025a) offer multi-metric evaluation but limited candidate generation. **(c)** MindDrive combines world-model-based trajectory generation with VLM-driven multi-objective reasoning for comprehensive selection.

In complex driving scenarios, human drivers reason in a structured, forward-looking manner: generating options, simulating their future evolution and risks, and selecting the optimal one through multi-objective trade-offs. Human-like autonomous driving similarly requires anticipating scene evolution and comparing behavioral options through structured reasoning. World models (Zheng et al., 2024a; 2025b; Li et al., 2024b) enable predictive scene imagination, while large language models provide complementary semantic and contextual reasoning for candidate evaluation (Bai et al., 2023; Touvron et al., 2023; Tang et al., 2025).

We propose **MindDrive**, a planning framework integrating foresighted trajectory generation and comprehensive selection. A **World Action Model (WAM)** performs what-if simulations to predict future scenarios and safety risks for each candidate, supporting a **Future-aware Trajectory Generator (FaTG)** that produces reliable, risk-sensitive trajectories. For selection, a learnable **VLM-oriented Evaluator (VLoE)** leverages VLM reasoning and generalization to assess candidates across multiple objectives and select the optimal plan.

Our contributions are as follows:

- We propose MindDrive, integrating high-quality trajectory generation and comprehensive decision reasoning in a harmonized framework.
- We introduce a **Future-aware Trajectory Generator (FaTG)** powered by a **World Action Model (WAM)** to simulate candidate trajectories' spatio-temporal evolution for foresighted planning.
- We design a **VLM-oriented Evaluator (VLoE)** that uses VLM semantic reasoning and generalization to assess trajectories across safety, comfort, and efficiency.
- Extensive experiments on NAVSIM-v1 and NAVSIM-v2 demonstrate strong performance improvements over state-of-the-art E2E-AD methods.

2 Related Work

2.1 End-to-End Autonomous Driving

As perception technologies mature, E2E-AD research increasingly focuses on planning. Think-Twice (Jia et al., 2023) highlights the importance of planning module design to overall performance. Recent advances follow three directions. The first strengthens prediction–planning coupling (Madjid et al., 2025): SparseDrive (Sun et al., 2025) jointly optimizes both tasks in a parallel architecture, while PPAD (Chen et al., 2024b) iteratively updates surrounding agents' intentions and ego plans to address limited feedback and game-theoretic interaction in cascaded pipelines. The second improves multimodal trajectory generation. VADv2 (Chen et al., 2024a) replaces single-output regression with a discrete vocabulary and sampling distribution. DiffusionDrive (Liao et al., 2025) generates diverse trajectories through diffusion and improves inference efficiency with truncated diffusion, while GoalFlow (Xing et al., 2025b) combines flow matching with goal-point mechanisms to improve trajectory quality and consistency. The third addresses temporal coherence: MomAD (Song et al., 2025a) uses trajectory and perception momentum for smoother, executable plans, while BridgeAD (Zhang et al., 2025) models historical and future contexts through multi-step temporal queries.

2.2 World Models for End-to-End Autonomous Driving

World models predict environmental evolution through temporal dynamics, allowing systems to rehearse future scenarios before execution. Existing approaches operate in three representation spaces. In 3D occupancy space, OccWorld (Zheng et al., 2024a) tokenizes semantic occupancy using VQ-VAE and performs temporal rollout with an autoregressive Transformer to decode ego trajectories. In BEV space, WoTE (Li et al., 2025b) predicts future BEV semantics conditioned on the current state and candidate trajectories, supporting online reward-based scoring and selection. In latent space, LAW (Li et al., 2024b) predicts next-step representations from scene features and ego actions, jointly learning latent features and trajectory prediction through self-supervision. World4Drive (Zheng et al., 2025b) forecasts candidate-specific latent states and selects trajectories by comparing predicted and actual/reconstructed latents. Think2Drive (Li et al., 2024a) uses a latent world model for transition learning, intrinsic rewards, and fast simulation in model-based reinforcement learning. Raw2Drive (Yang et al., 2025) extends it with a dual-stream model aligning privileged features and raw-sensor dynamics, improving training efficiency and generalization.

2.3 LLM-based Models for End-to-End Autonomous Driving

LLM-based models (Touvron et al., 2023; Liu et al., 2024; Wang et al., 2025b) enhance scene understanding, interpretability, and long-tail generalization through advanced reasoning. Their use in E2E-AD primarily takes two forms: **vision–language models (VLMs)** and **vision–language–action models (VLAs)**. VLM integration follows three patterns. Hierarchical designs, such as Senna (Jiang et al., 2024), feed VLM-generated meta-actions or semantic plans into an E2E planner for trajectory generation. LMDrive and related zero-shot LLM-assisted schemes provide periodic high-level instructions without closed-loop fine-tuning. Collaborative designs use VLMs as deliberative modules: FASIONAD++ (Qian et al., 2024) employs slow VLM reasoning to constrain a fast E2E planner and mitigate information bottlenecks, while DriveVLM (Tian et al., 2024) runs VLM and E2E branches in parallel for efficiency and safety redundancy. Holistic designs, such as ORION (Fu et al., 2025), align VLM semantic reasoning with the trajectory action space to condition trajectory generation. VLAs extend this integration by mapping vision and language directly to waypoints or control inputs. OpenEMMA (Xing et al., 2025a) uses an LLM policy to translate multimodal context into driving actions. OmniDrive (Wang et al., 2025a) integrates 3D scene modeling, language reasoning, and trajectory generation, using expert trajectories for supervised training and QA annotation. AutoVLA (Zhou et al., 2025) discretizes trajectories into action tokens to form candidate clusters and selects the optimal candidate using a reinforcement-learned policy score.

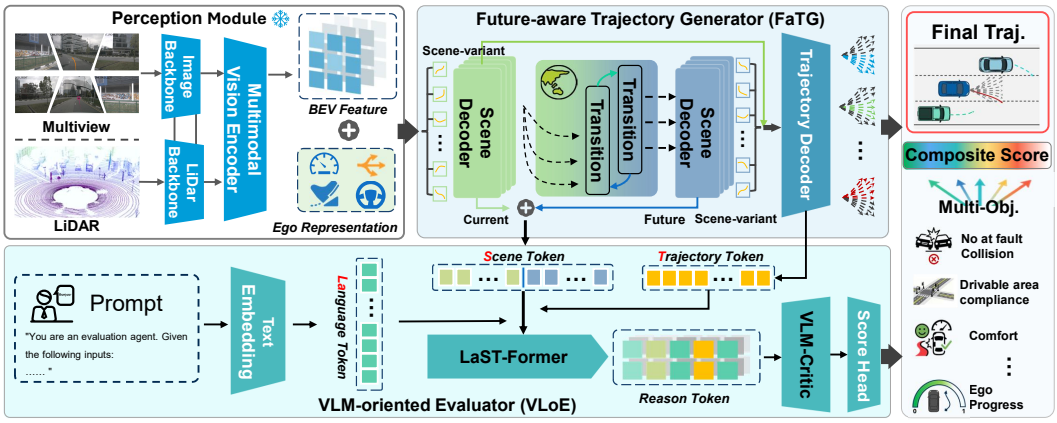


Figure 2: **Overview of the MindDrive framework.** The perception module fuses camera and LiDAR inputs into BEV features and encodes ego states and action intents as an **Ego Representation**. The **FaTG** constructs scene variants and predicts their evolution through **WAM**-based what-if simulations. The **VLoE** uses **LaST-Former** to fuse prompt and FaTG tokens for reasoning, and a **VLM-Critic** to score trajectories on safety, comfort, efficiency, and compliance, selecting the final trajectory by the aggregated score.

3 METHOD

3.1 OVERALL ARCHITECTURE

MindDrive (Fig. 2) integrates world modeling and VLM-based reasoning for trajectory generation and selection. Following TransFuser (Chitta et al., 2023), it fuses camera and LiDAR features into BEV features and encodes vehicle states and anchor-derived action intents as the ego representation. The **FaTG** uses **WAM** to predict future scenes under different ego intents, then combines current and predicted features through a trajectory decoder to generate candidates. The **VLoE** aligns language, scene, and trajectory tokens using **LaST-Former** for **VLM-Critic** evaluation, with a score head producing multi-objective scores for final trajectory selection.

3.2 FUTURE-AWARE TRAJECTORY GENERATOR

High-quality trajectory candidates support reliable decisions. However, prior methods typically rely on current-state-only predictions of surrounding agents, overlooking how the environment evolves under alternative ego decisions. To provide explicit foresight, the **Future-aware Trajectory Generator (FaTG)** combines a **World Action Model (WAM)** for ego-action-conditioned scene simulation with a **Trajectory Decoder** that generates candidates from current and predicted states.

3.2.1 WORLD ACTION MODEL

WAM (Fig. 3) simulates scene evolution under different ego intents through scene-variant construction and temporal rollout.

Scene-Variant Feature Construction. We derive N trajectory anchors by applying K-Means clustering (Chen et al., 2024a; Li et al., 2024c) to expert trajectories in the training set. A self-attention encoder processes each anchor, whose latent representation is concatenated with ego-status features and passed through an MLP to produce an action token $\mathbf{a}_{\text{token}}^{(n)}$ encoding the initial ego intent.

Let $\mathbf{B}_{\text{f}} \in \mathbb{R}^{H \times W \times C}$ denote the multimodal BEV feature map, with spatial dimensions $H \times W$ and C channels. Each action token is injected along its anchor trajectory using bilinear interpolation. Ego-centric coordinates (x, y) are mapped to BEV pixel coordinates as:

$$h = x\frac{H}{L_x}, \quad w = y\frac{W}{L_y} + \frac{W}{2}, \tag{1}$$

where L_x and L_y are the longitudinal and lateral BEV coverage in meters. The token is injected into the four neighboring pixels:

$$\mathbf{s}_{\mathrm{f}}^{(n)}(h_i, w_j) \leftarrow \mathbf{B}_{\mathrm{f}}(h_i, w_j) + w_{ij}\mathbf{a}_{\mathrm{token}}^{(n)}, \quad i, j \in \{0, 1\}, \tag{2}$$

where w_{ij} is the bilinear weight determined by the relative position of (h, w) within the neighboring grid points, and n indexes the action intent. This injection is summarized as:

$$\mathbf{s}_{\mathrm{f}}^{(n)} = \Phi(\mathbf{B}_{\mathrm{f}}, \mathbf{a}_{\mathrm{token}}^{(n)}), \tag{3}$$

where Φ denotes bilinear action-token injection. Stacking the resulting features yields:

$$\mathbf{s}_{\mathrm{f}} = [\mathbf{s}_{\mathrm{f}}^{(1)}, \ldots, \mathbf{s}_{\mathrm{f}}^{(N)}] \in \mathbb{R}^{N \times H \times W \times C}, \tag{4}$$

with each variant representing a scene conditioned on a distinct ego intent.

World-Model Rollout. WAM adopts a **spatial–temporal–spatial sandwich architecture**. A Transformer encoder first captures spatial interactions between the ego vehicle and its surroundings:

$$\mathbf{z}_t = \mathrm{Transformer}_{\mathrm{Enc}}(\mathbf{S}_{\mathrm{f},t}). \tag{5}$$

Temporal reasoning uses two Mamba blocks with pre-normalization and residual connections:

$$\begin{aligned} \mathbf{z}^{(1)} &= \mathbf{z}_t + \mathrm{Mamba}^{(1)}\left(\mathrm{PreNorm}^{(1)}(\mathbf{z}_t)\right), \\ \mathbf{z}^{(2)} &= \mathbf{z}^{(1)} + \mathrm{Mamba}^{(2)}\left(\mathrm{PreNorm}^{(2)}(\mathbf{z}^{(1)})\right), \\ \mathbf{z}_{t+k} &= \mathbf{z}^{(2)}. \end{aligned} \tag{6}$$

Here, (1) and (2) denote intermediate states from the two blocks. Mamba models temporal dependencies with linear computational complexity, while pre-normalization stabilizes training and residual connections preserve spatial information. Recurrent rollout predicts scene evolution over k time steps. A Transformer decoder then refines the evolved latent state and reconstructs future scene-variant features:

$$\mathbf{S}_{\mathrm{f},t+k} = \mathrm{Transformer}_{\mathrm{Dec}}(\mathbf{z}_{t+k}). \tag{7}$$

Repeated differentiable rollouts enable the planner to anticipate long-term outcomes for future-aware trajectory generation.

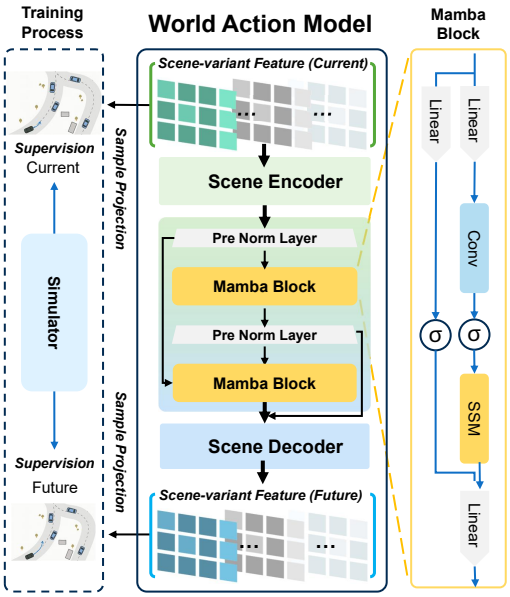


Figure 3: **World Action Model (WAM).** WAM adopts a **spatial–temporal–spatial sandwich** architecture, using Transformers for spatial encoding and reconstruction and Mamba for temporal rollout. Training is supervised by simulator-generated current and future BEV semantic map features.

3.2.2 Trajectory Decoder

The **Trajectory Decoder** combines ego action tokens with current and predicted scene features to generate trajectory candidates.

Scene-Feature Augmentation. For candidate n, current and future scene features are concatenated into a spatio-temporal embedding:

$$\boldsymbol{s}_{\mathrm{aug}}^{(n)} = \mathrm{Concat}\left(\boldsymbol{s}_{\mathrm{f},t}^{(n)}, \boldsymbol{s}_{\mathrm{f},t+k}^{(n)}\right). \tag{8}$$

Trajectory Candidate Generation. Each action token attends to the augmented features through multi-head cross-attention (MHCA). A decoding head converts the output into anchor offsets:

$$\boldsymbol{\tau}_n = \boldsymbol{\tau}_n^{\mathrm{anchor}} + \Psi_{\mathrm{offset}}\Big(\mathrm{MHCA}\big(\boldsymbol{a}_{\mathrm{token}}^{(n)}, \boldsymbol{s}_{\mathrm{aug}}^{(n)}\big)\Big). \tag{9}$$

The action token serves as the query and the augmented scene features as keys and values. The head Ψ_{offset} predicts offsets in $\mathbb{R}^{T \times d_{\mathrm{pose}}}$, producing future-aware candidates conditioned on the initial ego intent.

3.3 VLM-oriented Evaluator

Evaluating trajectory rationality and safety requires reasoning beyond geometric feasibility. The **VLM-oriented Evaluator (VLoE)** (Fig. 4) aligns language, scene, and trajectory representations for reasoning-driven assessment through two components: **LaST-Former** for multimodal fusion and **VLM-Critic** for language-guided evaluation.

3.3.1 LaST-Former

LaST-Former encodes scene and trajectory features into compact embeddings and integrates them with language cues through sentinel insertion, forming an aligned multimodal reasoning sequence.

Token Encoding. Current and future scene features are concatenated along the temporal dimension:

$$\mathbf{S}_{\text{token}} = \text{Concat}(\mathbf{S}_{\text{f},t}, \mathbf{S}_{\text{f},t+k}) . \tag{10}$$

A Conv2D network compresses spatial features into a compact embedding sequence, followed by a Linear–ReLU–Norm projection and multi-head cross-attention (MHCA):

$$\mathbf{S}_{\text{context}} = \text{MHCA}\Big(\text{Proj}\big(\text{Conv2D}(\mathbf{S}_{\text{token}}) \big) \Big), \tag{11}$$

where Proj is the projection head. This encoding captures spatial dynamics and contextual consistency across rollout steps.

Trajectory–Feature Encoding. The N candidate trajectories are concatenated into a token sequence:

$$\mathbf{T}_{\text{token}} = \text{Concat}(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_N) . \tag{12}$$

A Linear–ReLU–Norm projection maps these tokens into compact embeddings, which are refined through MHCA:

$$\mathbf{T}_{\text{context}} = \text{MHCA}\Big(\mathbf{Q}_{\text{metric}}, \mathbf{K} = \mathbf{V} = \text{Proj}(\mathbf{T}_{\text{token}}) \Big), \tag{13}$$

where $\mathbf{Q}_{\text{metric}} \in \mathbb{R}^{K \times C}$ contains learnable metric queries. These queries attend to all candidates, capturing motion intent and inter-trajectory dependencies for subsequent reasoning and scoring.

Sentinel Insertion. Scene and trajectory embeddings are serialized and inserted into the textual sequence using sentinel tokens such as `<scene>` and `<traj>`. During tokenization, sentinels receive reserved placeholder indices (e.g., -200), which are replaced by the corresponding multimodal embeddings at the embedding stage. This aligns scene, trajectory, and language features within a unified token sequence for VLM reasoning.

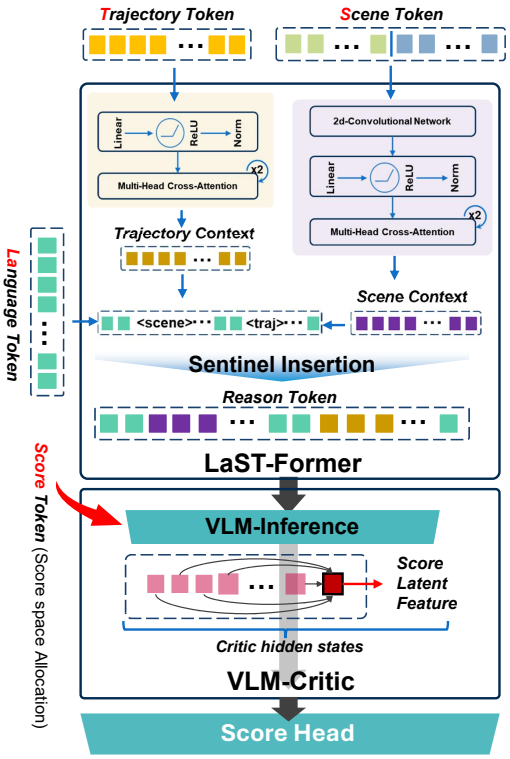


Figure 4: **VLM-oriented Evaluator (VLoE). LaST-Former** aligns and fuses prompt, trajectory, and scene tokens via sentinel insertion to produce unified **reasoning tokens**. The **VLM-Critic** uses a **score token** to aggregate scoring features and a score head to predict multi-objective trajectory scores.

3.3.2 VLM-Critic

We extend the base language model's vocabulary with a **score token** (`<score_feature>`) to aggregate textual and multimodal context for trajectory evaluation. The model processes the multimodal prompt causally, and the score token's hidden state serves as the evaluation feature:

$$\mathbf{h}_{\text{eval}} = \text{VLM}_{\text{infer}}(\mathbf{T}_{\text{reason}}, \mathbf{m}_{\text{attn}}) \Big|_{\texttt{score}}, \tag{14}$$

where $\mathbf{T}_{\text{reason}}$ contains the prompt and aligned scene–trajectory embeddings, $\mathbf{m}_{\text{attn}}$ controls token visibility, and $|_{\texttt{score}}$ extracts the score token's hidden state. A scoring head converts this feature into quantitative trajectory scores:

$$\mathbf{r}_{\text{score}} = \Psi_{\text{eval}}(\mathbf{h}_{\text{eval}}) , \tag{15}$$

based on predefined reasoning rules and assessment metrics.

4 EXPERIMENTS

4.1 DATASET AND METRICS

We evaluate on **NAVSIM** (Dauner et al., 2024; Cao et al., 2025), built on the nuPlan-derived OpenScene dataset (Contributors, 2023; Caesar et al., 2021), using the standard **Navtest** split and challenging subsets. **NAVSIM-v1** uses non-reactive simulation and **PDMS** to assess safety (NC, TTC), compliance (DAC), efficiency (EP), and comfort. Its **Navsafe** subset (Sima et al., 2025) covers ten safety-critical scenario categories mined using pre-crash typologies (Najm et al., 2007), human inspection, and CLIP-based clustering (Ilharco et al., 2021). **NAVSIM-v2** extends PDMS to **EPDMS**, adding direction compliance (DDC), traffic-light compliance (TLC), lane keeping (LK), and extended comfort (EC). Its **Navhard** split uses two-stage pseudo–closed-loop evaluation, combining scores from real observations and 3D Gaussian Splatting-generated future observations through Gaussian-weighted aggregation. Full metric names and descriptions are provided in Appendix A.1.

4.2 IMPLEMENTATION DETAILS

We adopt a vision–LiDAR fusion strategy to jointly leverage semantic and geometric cues from multimodal inputs. Synchronized front, front-left, and front-right RGB images are center-cropped, aligned, and concatenated into a 256×1024 composite, while LiDAR point clouds within a $64\,\mathrm{m} \times 64\,\mathrm{m}$ range around the ego vehicle are projected onto a BEV grid. A ResNet-34 (He et al., 2016) backbone encodes both modalities, whose features are extracted independently and fused via transformer-based cross-attention in BEV space, following TransFuser (Chitta et al., 2023). For high-level reasoning, we incorporate a Tiny-LLaVA-1B (Zhou et al., 2024) vision–language model as the inference backbone and fine-tune it using the LoRA strategy to efficiently adapt to driving-scene reasoning and evaluation tasks. The two-stage training procedure and scoring supervision are described in Appendix A.2.

4.3 MAIN RESULTS

We evaluate **MindDrive** under three NAVSIM protocols: v1-Navtest for overall planning quality, v2-Navtest for stricter compliance assessment, and v2-Navhard for robustness under synthetic future observations.

NAVSIM-v1 on Navtest. As shown in Table 1, MindDrive achieves state-of-the-art performance with a PDMS of **88.9**, outperforming DiffusionDrive (Liao et al., 2025), WoTE (Li et al., 2025b), and DIVER (Song et al., 2025b). It achieves strong performance across PDMS components, with strong safety (**TTC 95.1**), comfort (**Comf. 99.9**), and efficiency (**EP 82.8**).

NAVSIM-v2 on Navtest. We report only Stage I $EPDMS_1$, as Navtest lacks synthetic future observations for Stage II. Despite performance declines among baselines under the stricter criteria, MindDrive achieves an $EPDMS_1$ of **84.2**, with **99.3 DDC**, **96.5 LK**, and **96.8 EC**, demonstrating strong compliance and comfort (Table 2).

NAVSIM-v2 on Navhard. MindDrive achieves the highest EPDMS of **30.9** (Table 3). In Stage I, it achieves the highest DAC (**86.0**), with strong NC (**96.1**), TTC (**95.6**), and lane keeping (**LK 94.4**). Under Stage II synthetic observations, it remains competitive, achieving **79.1 DAC** and **98.0 TLC**, supporting its robustness under challenging future conditions.

4.4 ROBUSTNESS STUDY

We evaluate robustness on **Navsafe**, a safety-critical subset of Navtest. As shown in Table 4, MindDrive achieves the highest PDMS of **67.5** among reproduced baselines, with improvements in safety, compliance, efficiency, and comfort. These results support its planning robustness in rare, high-risk scenarios. Additional camera–LiDAR comparisons on Navhard under Stage I evaluation are provided in Appendix A.3.

Table 1: **NAVSIM-v1 results on Navtest.** C: Camera; L: LiDAR. All methods use a ResNet-34 backbone.

Method	Input	Venue	NC ↑	DAC ↑	EP ↑	TTC ↑	Comf. ↑	PDMS ↑
Human	-	-	100	100	87.5	100	99.9	94.8
LTF Chitta et al. (2023)	C	TPAMI 2022	97.4	92.8	79.0	92.4	100.0	83.8
TransFuser Chitta et al. (2023)	C & L	TPAMI 2022	97.7	92.8	79.2	92.8	**100**	84.0
UniAD Hu et al. (2023)	C	CVPR 2023	97.8	91.9	78.8	92.9	**100**	83.4
PARA-Drive Weng et al. (2024)	C	CVPR 2024	97.9	92.4	79.3	93.0	99.8	84.0
VADv2 Chen et al. (2024a)	C	arXiv 2024	97.9	91.7	77.6	92.9	**100**	83.0
DRAMA Yuan et al. (2024)	C & L	arXiv 2024	98.0	93.1	80.1	94.8	**100**	85.5
Hydra-MDP Li et al. (2024c)	C & L	arXiv 2024	98.3	96.0	78.7	94.6	**100**	86.5
Hydra-MDP++ Li et al. (2025a)	C & L	arXiv 2025	97.6	96.0	80.4	93.1	**100**	86.6
LAW Li et al. (2024b)	C	ICLR 2025	96.4	95.4	81.7	88.7	99.9	84.6
GoalFlow Xing et al. (2025b)	C & L	CVPR 2025	98.3	93.8	79.8	94.3	**100**	85.7
ARTEMIS Feng et al. (2025)	C & L	arXiv 2025	98.3	95.1	81.4	94.3	**100**	87.0
DiffusionDrive Liao et al. (2025)	C & L	CVPR 2025	98.2	96.2	82.2	94.7	**100**	88.1
WoTE Li et al. (2025b)	C & L	ICCV 2025	**98.5**	96.8	81.9	94.9	99.9	88.3
DIVER Song et al. (2025b)	C & L	arXiv 2025	**98.5**	96.5	82.6	94.9	**100**	88.3
MindDrive (Ours)	C & L	-	98.4	**97.0**	**82.8**	**95.1**	99.9	**88.9**

Table 2: **NAVSIM-v2 results on Navtest.** EPDMS_1 denotes the Stage I score; PDMS values match Table 1. All methods use a ResNet-34 backbone.

Method	NC↑	DAC↑	EP↑	TTC↑	HC↑	TLC↑	DDC↑	LK↑	EC↑	EPDMS_1 ↑	PDMS ↑
Transfuser Chitta et al. (2023)	97.7	92.8	79.2	92.8	100	99.9	98.3	67.6	95.3	77.8	84.0
VADv2 Chen et al. (2024a)	97.3	91.7	77.6	92.7	100	99.9	98.2	66.0	97.4	76.6	83.0
Hydra-MDP Li et al. (2024c)	97.5	96.3	80.1	93.0	100	99.9	98.3	65.5	97.4	79.8	86.5
Hydra-MDP++ Li et al. (2025a)	97.9	**96.5**	79.2	93.4	**100**	100	98.9	67.2	97.7	80.6	86.6
ARTEMIS Feng et al. (2025)	98.3	95.1	81.5	97.4	100	99.8	98.6	96.5	**98.3**	83.1	87.0
MindDrive (Ours)	**98.4**	95.6	**86.7**	**97.5**	**100**	99.8	**99.3**	**96.5**	96.8	**84.2**	**88.9**

Table 3: **NAVSIM-v2 results on Navhard.** EPDMS aggregates Stage I and Stage II scores with Gaussian weighting. All methods use a ResNet-34 backbone.

Method	Stage	NC↑	DAC↑	DDC↑	TLC↑	EP↑	TTC↑	LK↑	HC↑	EC↑	EPDMS↑
LTF Chitta et al. (2023)	Stage I	96.2	79.5	**99.1**	**99.5**	83.1	95.1	94.2	97.5	**79.1**	23.1
	Stage II	77.7	70.2	84.2	98.0	85.1	75.6	45.4	95.7	**75.9**	
DiffusionDrive Liao et al. (2025)	Stage I	96.0	79.7	97.4	**99.5**	81.3	93.1	90.8	96.8	73.8	24.2
	Stage II	82.1	72.2	88.5	98.7	85.1	78.8	49.2	89.3	71.2	
GTRS-DP Li et al. (2025c)	Stage I	94.7	78.8	96.1	**99.5**	83.0	94.4	92.0	97.5	72.8	23.8
	Stage II	80.3	74.4	84.9	98.0	81.9	78.8	45.4	96.7	70.1	
GuideFlow Liu et al. (2025)	Stage I	**96.6**	80.5	96.3	99.3	82.3	94.9	91.5	**97.7**	67.8	27.1
	Stage II	**87.3**	76.7	**88.8**	**99.2**	84.3	**85.1**	**49.7**	93.1	44.5	
MindDrive (Ours)	Stage I	96.1	**86.0**	98.8	99.3	**83.3**	**95.6**	**94.4**	97.6	74.7	**30.9**
	Stage II	82.6	**79.1**	86.4	98.0	**85.3**	79.4	49.2	96.5	**71.0**	

4.5 Ablation Study

Overall Architecture. We examine FaTG and VLoE on Navtest under NAVSIM-v1 (Table 5). FaTG improves safety and feasibility, while VLoE enhances overall PDMS, particularly EP and TTC. Combining both achieves the highest PDMS of **88.9**, supporting their complementary roles in future-aware generation and reasoning-guided selection.

WAM Architecture. Across four variants (Table 6), hybrid architectures outperform the pure Mamba model. The **Transformer–2×Mamba–Transformer** configuration performs best, supporting the combination of spatial attention and temporal modeling. An ablation of WAM temporal prediction steps is provided in Appendix A.4.

Table 4: **Robustness analysis on the Navsafe split under the NAVSIM-v1 metrics.** *: results reproduced using official weights.

Method	PDMS↑	NC↑	DAC↑	EP↑	TTC↑	Comf.↑
TransFuser* Chitta et al. (2023)	55.8	95.4	60.9	51.8	93.2	98.9
DiffusionDrive* Liao et al. (2025)	66.1	96.6	71.2	61.9	91.4	100
WoTE* Li et al. (2025b)	65.3	**98.7**	71.9	58.5	**96.3**	99.0
MindDrive	**67.5**	98.5	**75.6**	**63.1**	95.5	**100**

Table 5: **Ablation study of the overall architecture.** (On the Navtest split under the NAVSIM-v1 metrics.) FaTG: Future-aware Trajectory Generator. VLoE: VLM-oriented Evaluator.

FaTG	VLoE	PDMS ↑	NC ↑	DAC ↑	EP ↑	TTC ↑	Comf. ↑
×	×	84.1	97.8	92.5	79.3	93.0	99.8
✓	×	86.6	97.9	95.0	81.0	94.0	**99.9**
×	✓	87.7	98.1	96.4	81.7	94.2	**99.9**
✓	✓	**88.9**	**98.4**	**97.0**	**82.8**	**95.1**	**99.9**

Table 6: **WAM architecture ablation on Navtest (NAVSIM-v1).** Hybrid-Structure$^{1/2/3}$: 1: Transformer–Mamba; 2: Transformer–Mamba–Transformer; 3: Transformer–2×Mamba–Transformer.

Architecture	PDMS↑	NC↑	DAC↑	EP↑	TTC↑	Comf.↑
Mamba-Only	85.6	97.6	94.6	80.4	93.1	99.7
Hybrid-Structure^1	88.1	98.5	96.3	81.9	95.0	**99.9**
Hybrid-Structure^2	88.5	98.4	96.6	82.5	**95.1**	**99.9**
Hybrid-Structure^3	**88.9**	**98.4**	**97.0**	**82.8**	**95.1**	**99.9**

Table 7: **Ablation Study on VLoE.** (On the Navsafe split under the NAVSIM-v1 metrics.)

VLoE	PDMS↑	NC↑	DAC↑	EP↑	TTC↑	Comf.↑
Without VLM	64.5	97.6	61.0	59.4	93.1	99.7
With VLM	**67.5**	**98.5**	**75.6**	**63.1**	**95.5**	**100**

Table 8: **VLoE ablation on Navhard (NAVSIM-v2).** EPDMS_1: Stage I score without synthetic observations.

VLoE	EPDMS_1↑	NC↑	DAC↑	DDC↑	TLC↑	EP↑	TTC↑	LK↑	HC↑	EC↑
Without VLM	68.2	96.1	85.1	98.5	99.3	80.6	93.7	95.1	97.3	70.2
With VLM	**72.7**	**97.0**	**89.7**	**98.6**	**99.3**	**80.7**	**95.8**	**96.0**	**97.5**	**71.0**

VLoE Architecture. Adding VLM reasoning improves Navsafe PDMS from 64.5 to **67.5**, with gains in NC, DAC, and EP (Table 7), and Navhard $EPDMS_1$ from 68.2 to **72.7**, with gains in DAC, TTC, and LK (Table 8). These results support its benefits across both evaluation protocols. The Navhard ablation uses camera–LiDAR inputs and evaluates Stage I only, differing from the sensor configuration used in Table 3.

4.6 Visualization

Figure 5 compares MindDrive with TransFuser (Chitta et al., 2023) on challenging Navsafe scenarios. At intersections, MindDrive produces smoother, lane-consistent turns, while TransFuser exhibits trajectory drift and instability. In dense traffic, TransFuser deviates toward surrounding vehicles, risking collisions or near-misses, whereas MindDrive maintains stable, congestion-aware trajectories with safe distances and consistent driving intent.

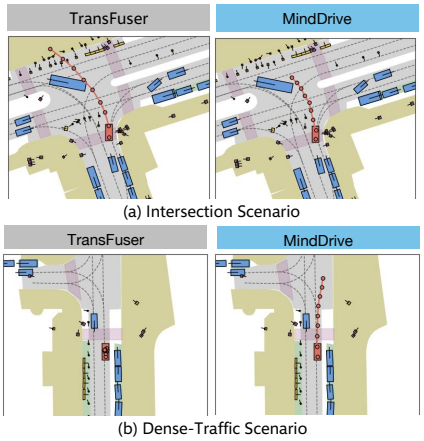


Figure 5: Qualitative comparison between TransFuser (Chitta et al., 2023) and our method on challenging Navsafe scenarios.

5 Conclusion

Our MindDrive introduces a structured reasoning paradigm that integrates what-if simulation, trajectory candidate generation, and multi-objective evaluation into a harmonized E2E-AD framework. The proposed Future-aware Trajectory Generator (FaTG), driven by a World Action Model (WAM), enables ego-conditioned "what-if" scene rollouts for foresighted trajectory planning. Building on these predictive candidates, the VLM-oriented Evaluator (VLoE) leverages a vision–language model to assess safety, comfort, and efficiency, yielding decisions that align with human driving intent. Extensive experiments on NAVSIM-v1 and NAVSIM-v2 demonstrate that MindDrive achieves consistent improvements in safety, smoothness, and generalization compared to state-of-the-art E2E-AD approaches. Future work includes incorporating reinforcement learning into the world–VLM reasoning loop to enable continuous self-improvement and further enhance system robustness in complex environments.

AI USE STATEMENT

In this work, we used generative AI tools to assist with mathematical formulation, method implementation, experimental analysis, and interpretation of results. Generative AI was not used to originate the core research problem, conceptual framework, or hypotheses. We did not use it to generate synthetic datasets, develop or write mathematical proofs, clean or reformat datasets, or conduct qualitative or thematic data analysis.

We also used generative AI to improve manuscript clarity and readability, assist with literature retrieval, and refine academic wording and the title. All AI-assisted content was reviewed and verified by the authors. AI-assisted code was tested, and mathematical formulations were checked against MindDrive's architecture, training objectives, and inference procedure. The authors take full responsibility for the final manuscript, experimental results, claims, and all AI-assisted materials.

REPRODUCIBILITY STATEMENT

The method section describes MindDrive, including the Future-aware Trajectory Generator (FaTG), World Action Model (WAM), and VLM-oriented Evaluator (VLoE). The experimental section specifies the benchmark datasets and splits, evaluation metrics, and experimental settings, and reports comparative results and component ablations. The appendix provides additional multimodal comparisons on Navhard and temporal prediction step analyses to support further evaluation of the reported results.

REFERENCES

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021.

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

Dian Chen and Philipp Krähenbühl. Learning from all vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17222–17231, 2022.

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*, 2024a.

Zhili Chen, Maosheng Ye, Shuangjie Xu, Tongyi Cao, and Qifeng Chen. Ppad: Iterative interactions of prediction and planning for end-to-end autonomous driving. In *European Conference on Computer Vision*, pp. 239–256. Springer, 2024b.

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.

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

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, et al. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *Advances in Neural Information Processing Systems*, 37:28706–28719, 2024.

Renju Feng, Ning Xi, Duanfeng Chu, Rukang Wang, Zejian Deng, Anzheng Wang, Liping Lu, Jinxiang Wang, and Yanjun Huang. Artemis: Autoregressive end-to-end trajectory planning with mixture of experts for autonomous driving. *arXiv preprint arXiv:2504.19580*, 2025.

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. *arXiv preprint arXiv:2503.19755*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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 (CVPR)*, pp. 17853–17862, June 2023.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, et al. Openclip zenodo, 2021.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21983–21994, June 2023.

Bo Jiang, Shaoyu Chen, Bencheng Liao, Xingyu Zhang, Wei Yin, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Senna: Bridging large vision-language models and end-to-end autonomous driving, 2024. URL `https://arxiv.org/abs/2410.22313`.

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.

Qifeng Li, Xiaosong Jia, Shaobo Wang, and Junchi Yan. Think2drive: Efficient reinforcement learning by thinking with latent world model for autonomous driving (in carla-v2). In *European Conference on Computer Vision*, pp. 142–158. Springer, 2024a.

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. *CoRR*, 2024b.

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*, 2025b.

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*, 2024c.

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*, 2025c.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12037–12047, 2025.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Lin Liu, Caiyan Jia, Guanyi Yu, Ziying Song, JunQiao Li, Feiyang Jia, Peiliang Wu, Xiaoshuai Hao, and Yandan Luo. Guideflow: Constraint-guided flow matching for planning in end-to-end autonomous driving. *arXiv preprint arXiv:2511.18729*, 2025.

Nadya Abdel Madjid, Abdulrahman Ahmad, Murad Mebrahtu, Yousef Babaa, Abdelmoamen Nasser, Sumbal Malik, Bilal Hassan, Naoufel Werghi, Jorge Dias, and Majid Khonji. Trajectory prediction for autonomous driving: Progress, limitations, and future directions. *arXiv preprint arXiv:2503.03262*, 2025.

Wassim G Najm, John D Smith, Mikio Yanagisawa, et al. Pre-crash scenario typology for crash avoidance research. Technical report, United States. Department of Transportation. National Highway Traffic Safety . . . , 2007.

Kangan Qian, Zhikun Ma, Yangfan He, Ziang Luo, Tianyu Shi, Tianze Zhu, Jiayin Li, Jianhui Wang, Ziyu Chen, Xiao He, et al. Fasionad: Fast and slow fusion thinking systems for human-like autonomous driving with adaptive feedback. *arXiv preprint arXiv:2411.18013*, 2024.

Chonghao Sima, Kashyap Chitta, Zhiding Yu, Shiyi Lan, Ping Luo, Andreas Geiger, Hongyang Li, and Jose M Alvarez. Centaur: Robust end-to-end autonomous driving with test-time training. *arXiv preprint arXiv:2503.11650*, 2025.

Ziying Song, Caiyan Jia, Lei Yang, Haiyue Wei, and Lin Liu. Graphalign++: An accurate feature alignment by graph matching for multi-modal 3d object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4):2619–2632, 2023.

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 Computer Vision and Pattern Recognition Conference*, pp. 22432–22441, 2025a.

Ziying Song, Lin Liu, Hongyu Pan, Bencheng Liao, Mingzhe Guo, Lei Yang, Yongchang Zhang, Shaoqing Xu, Caiyan Jia, and Yadan Luo. Breaking imitation bottlenecks: Reinforced diffusion powers diverse trajectory generation. *arXiv preprint arXiv:2507.04049*, 2025b.

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, 2025.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Pinxin Liu, Mingqian Feng, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2025. doi: 10.1109/TCSVT.2025.3566695.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22442–22452, 2025a.

Xin Wang, Yuwei Zhou, Bin Huang, Hong Chen, and Wenwu Zhu. Multi-modal generative ai: Multi-modal llms, diffusions and the unification. *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2025b. doi: 10.1109/TCSVT.2025.3635224.

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.

Shuo Xing, Chengyuan Qian, Yuping Wang, Hongyuan Hua, Kexin Tian, Yang Zhou, and Zhengzhong Tu. Openemma: Open-source multimodal model for end-to-end autonomous driving. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 1001–1009, 2025a.

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, 2025b.

Tianyu Yang, Yeqiang Qian, Weihao Yan, Chunxiang Wang, and Ming Yang. Adaptiveocc: Adaptive octree-based network for multi-camera 3d semantic occupancy prediction in autonomous driving. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

Zhenjie Yang, Xiaosong Jia, Qifeng Li, Xue Yang, Maoqing Yao, and Junchi Yan. Raw2drive: Reinforcement learning with aligned world models for end-to-end autonomous driving (in carla v2). *arXiv preprint arXiv:2505.16394*, 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 Computer Vision and Pattern Recognition Conference*, pp. 6854–6863, 2025.

Peiru Zheng, Yun Zhao, Zhan Gong, Hong Zhu, and Shaohua Wu. Simplevsf: Vlm-scoring fusion for trajectory prediction of end-to-end autonomous driving. *arXiv preprint arXiv:2510.17191*, 2025a.

Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. In *European conference on computer vision*, pp. 55–72. Springer, 2024a.

Wenzhao Zheng, Junjie Wu, Yao Zheng, Sicheng Zuo, Zixun Xie, Longchao Yang, Yong Pan, Zhihui Hao, Peng Jia, Xianpeng Lang, et al. Gaussianad: Gaussian-centric end-to-end autonomous driving. *arXiv preprint arXiv:2412.10371*, 2024b.

Yupeng Zheng, Pengxuan Yang, Zebin Xing, Qichao Zhang, Yuhang Zheng, Yinfeng Gao, Pengfei Li, Teng Zhang, Zhongpu Xia, Peng Jia, et al. 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, 2025b.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. Tinyllava: A framework of small-scale large multimodal models, 2024.

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.

A APPENDIX

A.1 EVALUATION METRICS

We summarize the metrics used in NAVSIM-v1 and NAVSIM-v2 (Dauner et al., 2024; Cao et al., 2025) below. All reported scores are higher-is-better.

Safety and Compliance. No-at-fault Collision (NC) measures avoidance of collisions attributable to the ego vehicle. **Time-to-Collision (TTC)** evaluates whether predicted collision risk remains within the benchmark's safety criteria; the reported value is a score rather than a time in seconds. **Drivable Area Compliance (DAC)** measures whether the vehicle remains within drivable areas. **Driving Direction Compliance (DDC)** assesses compliance with the prescribed driving direction, while **Traffic Light Compliance (TLC)** evaluates adherence to traffic signals.

Efficiency and Lane Keeping. Ego Progress (EP) measures progress along the route. **Lane Keeping (LK)** assesses whether the vehicle maintains an appropriate position within its lane.

Comfort. Comfort (Comf.) evaluates motion smoothness using kinematic criteria. **History Comfort (HC)** additionally accounts for consistency between the planned motion and the vehicle's recent motion history. **Extended Comfort (EC)** assesses consistency between successive plans over their overlapping prediction horizons.

Aggregate Scores. NAVSIM-v1 uses the **Predictive Driver Model Score (PDMS)** to combine NC, DAC, EP, TTC, and comfort. NAVSIM-v2 uses the **Extended Predictive Driver Model Score (EPDMS)**, incorporating NC, DAC, DDC, TLC, EP, TTC, LK, HC, and EC.

Two-Stage Evaluation. On Navhard, Stage I evaluates trajectories planned from real observations, while Stage II evaluates subsequent plans using synthetic future observations generated with 3D Gaussian Splatting around the Stage I endpoint. The final EPDMS combines both stages through Gaussian-weighted aggregation. In this paper, EPDMS_1 denotes the Stage I score when Stage II evaluation is not performed.

A.2 TRAINING PROCEDURE

MindDrive is trained in two stages. We first train the **World Action Model (WAM)** using current and future BEV semantic map features to learn scene evolution. We then jointly train WAM and the **VLM-oriented Evaluator (VLoE)** with trajectory generation and scoring objectives. For scoring supervision, the NAVSIM scorer computes metric-specific scores for 256 trajectory anchors, which serve as target labels for VLoE.

A.3 ADDITIONAL ROBUSTNESS EVALUATION

We further evaluate camera–LiDAR methods on Navhard under NAVSIM-v2 Stage I. Stage II uses camera-only synthetic observations and is therefore excluded from this multimodal comparison. The sensor configuration differs from that used in Table 3, so the corresponding Stage I scores are not directly comparable.

Table 9: **Robustness analysis on the Navhard split under the NAVSIM-v2 metrics.** *: denotes results reproduced from the official weights.† denotes methods operating on multimodal sensor inputs (camera + LiDAR). EPDMS_1 represents the EPDMS stage 1 evaluated without synthetic scenarios.

Method	EPDMS_1 ↑	NC↑	DAC↑	DDC↑	TLC↑	EP↑	TTC↑	LK↑	HC↑	EC↑
LTF* Chitta et al. (2023)	62.3	96.2	79.5	99.1	99.5	84.1	95.1	94.2	97.5	**79.1**
Transfuser† Chitta et al. (2023)	60.5	96.3	74.6	98.4	99.3	82.9	93.7	92.7	97.5	78.2
DiffusionDrive† Liao et al. (2025)	63.2	95.9	84.0	**98.6**	**99.8**	**84.4**	96.0	95.1	**97.6**	71.1
WoTE† Li et al. (2025b)	66.7	**97.4**	88.2	97.8	99.3	82.7	**96.4**	90.9	97.3	68.0
MindDrive†	**72.7**	97.0	**89.7**	**98.6**	99.3	80.7	95.8	**96.0**	97.5	71.0

As shown in Table 9, MindDrive achieves the highest $EPDMS_1$ of **72.7**, with strong safety and compliance scores: **97.0 NC**, **89.7 DAC**, **98.6 DDC**, and **95.8 TTC**. These results support its robustness in challenging scenarios with multimodal inputs.

A.4 Temporal Prediction Step Analysis

We investigate the temporal prediction strategy of WAM by comparing direct long-horizon prediction (0s $\rightarrow$ 4s) with recurrent short-step prediction (0s $\rightarrow$ 2s $\rightarrow$ 4s). Here, $A \rightarrow B$ denotes predicting the scene state at time B from the state at time A.

Table 10: **Ablation Study on World Action Model time steps.** (On the Navtest split under the NAVSIM-v1 metrics.)

Rollout Steps	PDMS ↑	NC ↑	DAC ↑	EP ↑	TTC ↑	Comf. ↑
0s $\rightarrow$ 4s	87.5	98.1	96.1	81.7	94.3	**99.9**
0s $\rightarrow$ 2s $\rightarrow$ 4s	**88.9**	**98.4**	**97.0**	**82.8**	**95.1**	**99.9**

As shown in Table 10, recurrent prediction outperforms direct prediction across NC, DAC, EP, TTC, and PDMS. These results support incorporating intermediate predicted states into world-model rollout for future-aware trajectory planning.