REWAM: RECIPROCAL WORLD ACTION MODELS FOR INTERACTIVE AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

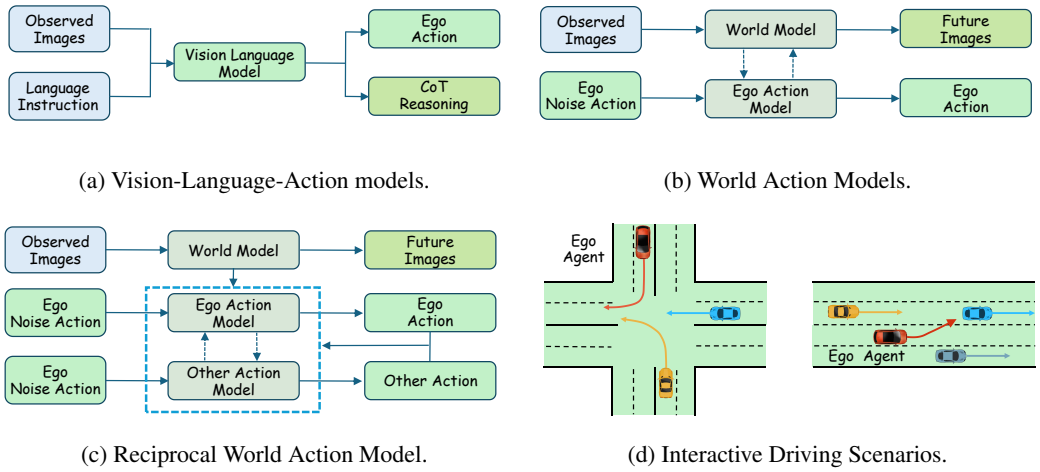
In interactive scenarios, an autonomous driving system is required to generate ego actions under the influence of other agents' behaviors. Existing World Action Models (WAMs) typically model other agents as components of the world model rather than as decision-makers that fundamentally shape the action of the ego agent, which impairs their performance in dense interaction scenarios. We introduce Reciprocal World Action Models (**ReWAM**), a game-theoretic world action modeling framework that captures the reciprocal influence between the ego agent and other agents by representing them as conditional responders whose actions are mutually influenced. We instantiate this framework with a Level-k response hierarchy, where role-specific ego and other action DiTs exchange compact strategy tokens through cross-agent attention while remaining grounded in a shared representation of the future driving world. To learn the response policy of the ego agent from demonstrations, we formulate expert actions as samples from the best response distribution and jointly optimize the entire hierarchy using conditional flow matching. Our framework is evaluated on the NAVSIM dataset and achieves state-of-the-art performance compared to baselines. The improvement is particularly significant in interactive scenarios, validating that modeling reciprocal responses provides a more effective foundation for interaction-aware world action generation.

1 INTRODUCTION

Autonomous driving requires an intelligent system to perceive complex traffic scenarios, anticipate their evolution, and plan a safe policy while interacting with multiple road users whose decisions are mutually dependent. End-to-end (Wang et al., 2026b; Guo et al., 2026; Xing et al., 2025; Jia et al., 2025b; Shang et al., 2026) autonomous driving simplifies conventional modular pipelines by jointly learning driving representations and action directly from historical sensor observations. Recent methods have improved this paradigm through multimodal feature fusion, goal-conditioned action generation, unified task modeling, and preference-aligned policy learning. However, the end-to-end paradigm primarily learns a direct mapping from observations to action, making it effective at reproducing demonstrated behavior, but action supervision alone does not ensure the policy acquires an explicit understanding of the environment and the mechanisms governing its evolution, thereby limiting its ability to generalize to real-world scenarios that fall outside the training distribution.

Compared with direct observation-to-action policies, Vision-Language-Action (VLA) models introduce language as a semantic interpreter between visual perception and action generation. By aligning visual observations with linguistic concepts, they are capable of interpreting road semantics, driving rules, and traffic participants' behaviors, and reasoning about an appropriate driving decision (Jiang et al., 2025; Kim et al., 2024; Wang et al., 2025; Zhou et al., 2026a; Liu et al., 2025; Zhou et al., 2026c). Reasoning-centric extensions further employ structured reasoning, cognitive feedback, and self-improvement to connect language-based scene understanding with executable action (Liu et al., 2026; Xiong et al., 2026; Zhang et al., 2026; Xiao et al., 2026). More recent methods combine VLA policies with prospective scene prediction or latent world modeling, allowing decisions to exploit imagined futures (Cen et al., 2025; Jia et al., 2026; Wang et al., 2026a; Fan et al., 2026; Xie et al., 2026; Chen et al., 2026a). These advances substantially strengthen semantic scene understanding and high-level reasoning. However, reasoning in the language space alone does not explicitly model the continuous physical evolution of the environment under different actions.

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.



(a) Vision-Language-Action models.

(b) World Action Models.

(c) Reciprocal World Action Model.

(d) Interactive Driving Scenarios.

Figure 1: Comparison of different autonomous driving paradigms and representative interactive driving scenarios. (a) Vision-Language-Action models incorporate language-aligned semantic reasoning into action generation. (b) World Action Models ground ego-action generation in predicted world representations. (c) The proposed Reciprocal World Action Model explicitly captures reciprocal behavioral influence through recursive strategic reasoning. (d) Representative interactive scenarios in which the behaviors of the ego agent and other agents are mutually influenced.

A semantically plausible decision may therefore lack a grounded prediction of its consequences in the future world.

World models address a complementary limitation by learning how driving scenarios evolve over time. They forecast future observations or latent states conditioned on scenario history and candidate action, enabling a policy to evaluate possible consequences before acting (Zhou et al., 2026b; Zheng et al., 2025; Peng et al., 2026; Ye et al., 2026b). World Action Models (WAMs) further couple future generation and action prediction within a shared architecture, exposing predictive world representations directly to the planning process (Xia et al., 2026; Chen et al., 2026b; Wang et al., 2026c; Yuan et al., 2026; Ye et al., 2026a; Maes et al., 2026). Many predictive-representation approaches have demonstrated the value of compact latent dynamics for action-conditioned reasoning (Yang et al., 2026; Sun et al., 2026; LeCun et al., 2022). Despite this progress, existing WAMs primarily model how the ego action influences the future world. Other agents are usually predicted as part of that world, but their actions are not explicitly organized as strategic responses to the ego plan. Consequently, an action-conditioned future can be visually plausible without being behaviorally consistent with the interaction induced by other agents.

Interaction-aware prediction and planning methods explicitly model this reciprocal influence through conditional trajectory prediction, joint planning, or iterative behavioral refinement (Huang et al., 2023; Jia et al., 2025a; Tan et al., 2026; Li et al., 2025a; Liang et al., 2025). However, these methods predominantly operate in explicit state, maneuver, or action spaces, and their interaction modules are often separated from generative world modeling. This reveals a gap between two complementary directions: WAMs learn expressive action-conditioned representations of future driving worlds but lack an explicit strategic reasoning structure, whereas interaction-aware prediction and planning model behavioral responses but do not directly exploit the latent representations learned by generative world models.

Bridging these two directions requires more than adding an auxiliary other-agent prediction head to a WAM. A suitable interaction mechanism should preserve the learned representation of future world dynamics, represent ego and other agents' decisions as mutually dependent, and remain tractable in scenes involving multiple agents. In particular, it should capture a finite chain of behavioral responses without requiring the computation of a fully rational multi-agent equilibrium. A level-k game is a non-equilibrium model of bounded rationality that organizes agents into a hierarchy of strategic types, making it particularly well suited to autonomous driving. To date, the level-k game has been applied to lane changing, speed management, unsignalized intersections, and contingency

planning (Karimi et al., 2023; Yuan et al., 2023; Chen et al., 2026c; Fang & Yi, 2024). However, existing formulations generally operate over handcrafted states, discrete maneuvers, or explicit actions. To the best of our knowledge, integrating the Level-k strategic hierarchy directly into the latent representations of a generative WAM remains unexplored.

In this paper, we propose **ReWAM**, a game-theoretic framework that incorporates bounded strategic reasoning into world-action modeling to explicitly capture how the ego agent and other agents mutually influence one another's actions. Our core idea is to represent the ego agent and other agents as mutually responsive decision makers grounded in a shared future-world representation. At each reasoning level, the ego and other agents' action branches synchronously update their action hypotheses by conditioning on the joint strategy predicted at the preceding level. This design retains the expressive latent dynamics of a WAM while making reciprocal behavioral responses explicit in action generation. Our contributions are summarized as follows:

- We explicitly model the reciprocal influence between ego and other agents' actions within a WAM, extending WAMs beyond one-way ego-action conditioning toward interaction-aware action generation.
- We propose ReWAM, a game-theoretic world-action modeling framework that integrates bounded reasoning with role-specific action DiTs, enabling the consideration of the reciprocal influence between the ego and other agents' actions.
- We evaluate the proposed ReWAM on NAVSIM, where it achieves state-of-the-art (SOTA) performance and delivers particularly pronounced improvements in scenarios with dense interactions, which establishes reciprocal response modeling as a critical capability for robust world-action generation in interactive scenarios.

2 METHODOLOGY

In this section, we introduce the proposed ReWAM framework, a game-theoretic WAM that augments generative world grounding with finite-depth strategic reasoning for interactive autonomous driving. Rather than treating the behavior of other agents as an implicit component of future-world prediction, ReWAM casts ego and other agents' actions as mutually conditioned responses within a shared latent representation of the world. As illustrated in Figure 2, we formalize this hierarchy as a finite Level-k game, instantiate its response policies with separate ego and other-agent action models grounded in a shared world model, and learn the resulting conditional responses from expert demonstrations.

2.1 LEVEL-k GAME FOR SOLVING INTERACTIVE POLICY

Consider a scenario containing an ego agent e and N other agents, with participant set $\mathcal{P} = \{e, 1, \ldots, N\}$. At time t, the model receives an observation $\mathcal{O}_t$ containing the visual history, textual navigation condition, and current participant states s_t. Let τ_i denote the future action of participant i, and let $\tau_{-i} = (\tau_j)_{j \in \mathcal{P} \setminus \{i\}}$ indicate the joint action of all remaining participants. Similarly, π_i denotes the policy of participant i, and π_{-i} represents the corresponding joint policy of the remaining participants. Given π_{-i}, the expected return and best-response set of participant i are defined as

$$J_i(\pi_i, \pi_{-i}; \mathcal{O}_t) = \mathbb{E}_{\boldsymbol{\tau} \sim (\pi_i, \pi_{-i})} \left[R_i(\boldsymbol{\tau}; \mathcal{O}_t) \right], \tag{1}$$

$$\mathrm{BR}_i(\pi_{-i}; \mathcal{O}_t) = \arg\max_{\pi_i} J_i(\pi_i, \pi_{-i}; \mathcal{O}_t), \tag{2}$$

where $\boldsymbol{\tau} = (\tau_i, \tau_{-i})$ is the joint action and R_i is the reward of participant i. Directly coupling all participants at the same reasoning level produces circular dependencies and would generally require solving for a joint equilibrium, which is computationally prohibitive for real-time autonomous driving. Level-k game instead replaces such simultaneous coupling with a hierarchy of bounded strategic responses:

$$\begin{aligned} \pi_i^{(k)} &\in \mathrm{BR}_i\left(\pi_{-i}^{(k-1)}; \mathcal{O}_t\right), \\ \mathrm{BR}_i\left(\pi_{-i}^{(k-1)}; \mathcal{O}_t\right) &= \arg\max_{\pi_i} \mathbb{E}_{\boldsymbol{\tau} \sim (\pi_i, \pi_{-i}^{(k-1)})} \left[R_i(\boldsymbol{\tau}; \mathcal{O}_t) \right], \qquad k \geq 1. \end{aligned} \tag{3}$$

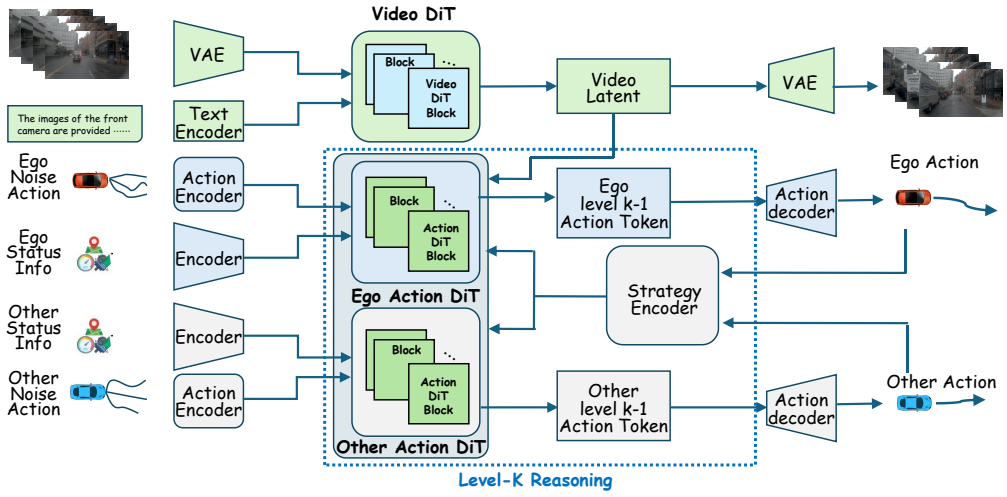


Figure 2: The proposed ReWAM framework. Visual observations and language instructions are processed by the video branch to construct a shared latent representation of the future driving world. Conditioned on this representation and agent-specific state information, the Ego and Other Action DiTs synchronously generate their action hypotheses. At reasoning level k, each branch conditions on the level $k-1$ strategies of all remaining valid participants, forming a finite hierarchy of reciprocal responses.

where $\pi_i^{(k)}$ denotes participant i's policy at reasoning level k. This formulation converts simultaneous mutual dependence into a finite sequence of strategic responses, preserving reciprocal interaction while avoiding the computation of a joint equilibrium.

2.2 World Representation Conditioned Level-k Reasoning

As illustrated in Figure 2, ReWAM embeds the strategic hierarchy described in equation 3 into a generative WAM architecture, where a shared world model provides predictive context and dedicated Action DiTs parameterize the response policies of the ego agent and other agents. The world model utilizes the visual observations and textual navigation condition to generate a shared latent representation of the future driving world:

$$\mathbf{W}_t = f_\phi^{\mathrm{WM}}(I_{t-H:t}, c_t) \in \mathbb{R}^{L_v \times d_v}, \tag{4}$$

where $I_{t-H:t}$ denotes the sequence of historical visual observations up to time t, c_t denotes the textual navigation condition, and L_v and d_v denote the number and dimension of world tokens, respectively. The same world tokens $\mathbf{W}_t$ are provided to all action branches so that their responses are grounded in a consistent representation of the driving world.

Grounded in the shared world representation, ReWAM employs two role-specific response generators: an Ego Action DiT parameterized by θ_e, which predicts the ego agent's response to the preceding-level strategies of all other agents, and an Other Action DiT parameterized by θ_o, which is shared across other agents and predicts each other agent's response. Let $\mathbf{Z}_{-i}^{(k-1)}$ denote the strategy encoder representation of the level $k-1$ action of all participants other than i. Conditioned on a realized preceding-level strategy, the current action hypotheses are sampled as

$$\tau_e^{(k)} \sim q_{\theta_e}\left(\cdot \mid \mathbf{W}_t, s_t^e, \mathbf{Z}_{-e}^{(k-1)}\right), \tag{5}$$

$$\tau_i^{(k)} \sim q_{\theta_o}\left(\cdot \mid \mathbf{W}_t, s_t^i, \mathbf{Z}_{-i}^{(k-1)}\right), \qquad i = 1, \ldots, N \tag{6}$$

where $(\mathbf{W}_t, s_t^i)$ represents the decision-relevant information in $\mathcal{O}_t$ for participant i. Therefore, equation 5 and equation 6 provide amortized, action-conditioned realizations of the bounded responses in equation 3, where every level k prediction responds to a fixed realization of the remaining participants' level $k-1$ strategies, rather than solving a simultaneous equilibrium online.

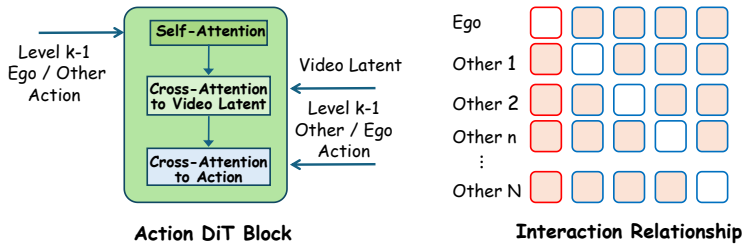


Figure 3: Architecture of the proposed Level-k interaction block. Within each Action DiT block, temporal self-attention models the target agent's action sequence, while cross-attention to the Video DiT features grounds action generation in the predicted driving world. Cross-agent attention further conditions the current response on the preceding-level action hypotheses of the remaining agents. The elaborated Action DiT architecture explicitly captures reciprocal strategic influence among agents in interactive scenarios.

2.3 Learning Best Responses from Demonstrations

Since specifying driving rewards for real-world scenarios is rather challenging, it is difficult to directly compute the best-response policies characterized in equation 2 . We instead assume that the demonstrated action is generated by a joint expert policy $\pi^{\mathrm{E}} = (\pi_j^{\mathrm{E}})_{j\in\mathcal{P}}$, whose components are reward-maximizing responses to the remaining expert policies:

$$\pi_i^{\mathrm{E}} \in \mathrm{BR}_i\left(\pi_{-i}^{\mathrm{E}}; \mathcal{O}_t\right), \tag{7}$$

$$\boldsymbol{\tau}^{\mathrm{demo}} \sim \pi^{\mathrm{E}}(\cdot \mid \mathcal{O}_t), \tag{8}$$

where $\boldsymbol{\tau}^{\mathrm{demo}}$ is the demonstrated joint action. Let $p_{\mathrm{demo}}(\tau_i \mid \mathcal{O}_t, \tau_{-i})$ denote the conditional response distribution induced by the joint demonstrations. After applying the world, state, and strategy encoders, we use the same notation for the corresponding model-level conditional distribution $p_{\mathrm{demo}}(\tau_i \mid \mathbf{W}_t, s_t^i, \mathbf{Z}_{-i})$. The response model can be learned through conditional distribution matching:

$$\theta_i^\star = \arg\min_{\theta_i} \mathcal{D}\left(p_{\mathrm{demo}}\left(\tau_i \mid \mathbf{W}_t, s_t^i, \mathbf{Z}_{-i}\right), q_{\theta_i}\left(\tau_i \mid \mathbf{W}_t, s_t^i, \mathbf{Z}_{-i}\right)\right), \tag{9}$$

where $\theta_i = \theta_e$ for the ego vehicle and $\theta_i = \theta_o$ for other agents, respectively; and $\mathcal{D}$ measures the discrepancy between the demonstrated and learned conditional distributions. The resulting q_{θ_i} is therefore an amortized, stochastic surrogate for the best-response policy over the support of the demonstrated strategy distribution.

2.4 Level-k Interaction Block

As illustrated in Figure 2, ReWAM parameterizes the response distributions of the ego agent and other agents using two Action DiTs that progressively transform noisy action sequences into world-grounded and interaction-conditioned action predictions through stacked transformer blocks. Each block successively applies temporal self-attention over the target participant's future action tokens, and cross-attention to the corresponding Video DiT features, cross-agent attention to the preceding-level action hypotheses, and a feed-forward transformation. This architecture integrates temporal action modeling, world-grounded contextualization, and explicit strategic interaction within one WAM framework. As shown in Figure 3, to model the interaction among agents, the cross-attention for participant i at transformer block ℓ is computed as

$$\mathbf{A}_{i,\ell}^{(k)} = \mathrm{softmax}\left(\frac{Q_\ell(\mathbf{H}_{i,\ell}^{(k)})K_\ell(\mathbf{Z}^{(k-1)})^\top}{\sqrt{d_h}} + \mathbf{M}_i\right) V_\ell(\mathbf{Z}^{(k-1)}), \tag{10}$$

where $\mathbf{H}_{i,\ell}^{(k)} \in \mathbb{R}^{T\times d}$ denotes the action tokens of participant i; d_h is the attention-head dimension; the target participant's action tokens serve as queries Q_ℓ, while the strategy representations of the preceding-level action provide keys K_ℓ and values V_ℓ. The mask $\mathbf{M}_i$ imposes role-dependent visibility by suppressing padded participants and the target participant's own strategy token. Therefore, the proposed interaction blocks enable cross-agent attention to explicitly model the mutual influence among participants by conditioning each participant's current response on the preceding-level action hypotheses of all other valid participants.

2.5 JOINT LEARNING AND INFERENCE

Flow matching provides a tractable velocity-regression formulation of the distribution matching objective, so it is adopted to learn the response model targeted by equation 9. For participant $j \in \mathcal{P}$ at reasoning level k, let $\mathbf{x}_{0,j}$ denote a normalized demonstrated action sampled from the target conditional distribution:

$$\mathbf{x}_{0,j} \sim p_{\text{demo}}\Big(\cdot \mid \mathbf{W}_t, s_t^j, \mathbf{Z}_{-j}^{(k-1)}\Big). \tag{11}$$

Then, we sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $r \sim \mathcal{U}(0,1)$, and define

$$\mathbf{x}_{r,j} = (1-r)\mathbf{x}_{0,j} + r\boldsymbol{\epsilon}, \tag{12}$$

$$\mathbf{u}_{r,j} = \boldsymbol{\epsilon} - \mathbf{x}_{0,j}. \tag{13}$$

where $\boldsymbol{\epsilon}$ indicates a Gaussian noise sample that defines the source distribution; r is the continuous flow time that interpolates between the demonstrated action and noise; $\mathbf{x}_{r,j}$ denotes the action state at flow time r along the linear interpolation between the demonstrated action and Gaussian noise; $\mathbf{u}_{r,j}$ represents the corresponding target velocity. The resulting state-velocity pairs provide the regression targets for the conditional velocity field, which is learned by minimizing

$$\mathcal{L}_j^{(k)} = \mathbb{E}\left[\left\| v_{\theta_j}\Big(\mathbf{x}_{r,j}, r \mid \mathbf{W}_t, s_t^j, \mathbf{Z}_{-j}^{(k-1)}\Big) - \mathbf{u}_{r,j} \right\|_2^2\right]. \tag{14}$$

Equation 14 defines the participant-level flow-matching objective for a single response at reasoning level k, and these objectives are subsequently aggregated according to the two response roles:

$$\mathcal{L}_{\text{ego}}^{(k)} = \mathcal{L}_j^{(k)}\Big|_{j=e}, \qquad \mathcal{L}_{\text{other}}^{(k)} = \frac{\sum_{i=1}^N m_i \mathcal{L}_i^{(k)}}{\sum_{i=1}^N m_i}, \tag{15}$$

where $m_i \in \{0,1\}$ indicates the validity of surrounding agent i. The branch-level objectives are aggregated across the reasoning hierarchy using normalized depth-dependent weights:

$$\mathcal{L}_{\text{flow}} = \sum_{k=1}^{K_{\text{train}}} \alpha_k \left(\lambda_{\text{ego}} \mathcal{L}_{\text{ego}}^{(k)} + \lambda_{\text{other}} \mathcal{L}_{\text{other}}^{(k)}\right), \qquad \alpha_k = \frac{k^p}{\sum_{\ell=1}^{K_{\text{train}}} \ell^p}. \tag{16}$$

where K_{train} specifies the supervised reasoning depth; α_k denotes the normalized depth-dependent weight; p controls the relative emphasis assigned to deeper response levels; λ_{ego} and λ_{other} balance the two action branches; the normalization ensures $\sum_{k=1}^{K_{\text{train}}} \alpha_k = 1$. Accordingly, the response models are jointly optimized under the aggregate objective:

$$\Theta^\star \in \arg\min_{\Theta} \mathcal{L}_{\text{flow}}(\Theta), \tag{17}$$

where Θ denotes the complete set of trainable response-model parameters. At the population optimum, conditional flow matching recovers the demonstration-conditioned response distribution for each valid participant and supervised reasoning level (see Appendix A.2), which means

$$q_{\theta_j^\star}\Big(\cdot \mid \mathbf{W}_t, s_t^j, \mathbf{Z}_{-j}^{(k-1)}\Big) = p_{\text{demo}}\Big(\cdot \mid \mathbf{W}_t, s_t^j, \mathbf{Z}_{-j}^{(k-1)}\Big), \qquad j \in \mathcal{P}, \quad k = 1, \ldots, K_{\text{train}}. \tag{18}$$

where $\theta_j^\star = \theta_e^\star$ for the ego agent and $\theta_j^\star = \theta_o^\star$ for other agents. Therefore, the distribution-matching objective in equation 9 is achieved across response roles and reasoning levels by conditional flow matching.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Implementation details. We implement the proposed ReWAM on top of DriveLaW (Xia et al., 2026) by designing the Level-k interaction block and introducing a parameter-shared Other Action DiT together with a strategy encoder. The pretrained VAE, text encoder, and Video DiT are frozen, while both Action DiTs and the associated interaction modules are optimized. We first train the Level-1 model and initialize every $k \geq 2$ model from the resulting checkpoint for further fine-tuning. We use four front-camera frames sampled at 2 Hz and predict eight poses at 0.5 s intervals for the ego agent and at most ten other agents within 80 m. Current agent states are obtained from scene annotations, whereas future agent states and actions are used only as supervision.

Table 1: Comparison with state-of-the-art methods on the NAVSIM benchmark. Higher values indicate better performance.

Method	Ref	NC↑	DAC↑	TTC↑	Comf.↑	EP↑	PDMS↑
Traditional End-to-End Methods							
VADv2-V_{8192} (Jiang et al., 2024)	arXiv'24	97.2	89.1	91.6	**100**	76.0	80.9
UniAD (Hu et al., 2023)	CVPR'23	97.8	91.9	92.9	**100**	78.8	83.4
TransFuser (Chitta et al., 2023)	TPAMI'23	97.7	92.8	92.8	**100**	79.2	84.0
PARA-Drive (Weng et al., 2024)	CVPR'24	97.9	92.4	93.0	99.8	79.3	84.0
DiffusionDrive (Liao et al., 2025)	CVPR'25	98.2	96.2	94.7	**100**	82.2	88.1
ReCogDrive-IL (Xiong et al., 2026)	ICLR'26	98.1	94.7	94.2	**100**	80.9	86.5
World Model Methods							
DrivingGPT (Chen et al., 2025)	ICCV'25	98.9	90.7	94.9	95.6	79.7	82.4
LAW (Li et al., 2025b)	ICLR'25	96.4	95.4	88.7	99.9	81.7	84.6
Epona (Zhang et al., 2025)	ICCV'25	97.9	95.1	93.8	99.9	80.4	86.2
WoTE (Li et al., 2025c)	ICCV'25	98.5	96.8	94.9	99.9	81.9	88.3
DriveVLA-W0 (Li et al., 2026)	ICLR'26	98.4	95.3	95.2	**100**	80.9	87.2
PWM (Zhao et al., 2025)	NeurIPS'25	98.6	95.9	95.4	**100**	81.8	88.1
WorldDrive (Gui et al., 2026)	arXiv'26	98.4	96.8	95.2	**100**	83.3	89.0
DriveLaW (Xia et al., 2026)	CVPR'26	99.0	97.1	96.7	**100**	81.3	89.1
ReWorld (Peng et al., 2026)	arXiv'26	99.1	**98.2**	**97.7**	**100**	82.0	90.4
ReWAM (Ours)	-	**99.3**	97.8	97.2	**100**	**83.6**	**90.6**

Dataset and evaluation metrics. We conduct experiments on NAVSIM, a planning-oriented benchmark derived from the public OpenScene/nuPlan driving logs. We use the official `navtrain` split for training and the `navtest` split for evaluation. The latter contains 136 logs and 12,146 driving scenarios. Unless stated otherwise, we report the Predictive Driver Model Score (PDMS) together with its components: No-at-Fault Collision (NC), Drivable Area Compliance (DAC), Time to Collision (TTC), Comfort, and Ego Progress (EP). We additionally report Driving Direction Compliance (DDC) as a diagnostic measure. All metric values are multiplied by 100, and higher values indicate better performance.

3.2 Main Results

Comparison with SOTA methods. As presented in Table 1, ReWAM achieves the highest PDMS score of 90.6, improving over its DriveLaW backbone by 1.5 points and the SOTA method, ReWorld, by 0.2 points. The gain over DriveLaW is accompanied by improvements in NC, DAC, TTC, and EP scores, with the largest enhancement appearing in EP score. Although ReWorld obtains slightly higher DAC and TTC scores, ReWAM achieves better performance in NC and EP scores and consequently a better aggregate score. These results validate that explicitly modeling the interaction between the ego agent and other agents enables the planner to better anticipate reciprocal behavioral responses, thereby achieving a SOTA performance on the benchmark.

3.3 Analysis of Level-k Reasoning

Effect of reasoning depth We vary the reasoning depth while keeping the model and evaluation protocol fixed, and report the performance at each reasoning level in Table 2. Increasing the depth from Level-1 to Level-4 consistently improves the aggregate performance, raising PDMS from 90.03 to 90.64. The largest single-level gain occurs at Level-2, whereas subsequent levels deliver progressively smaller improvements. This trend is most pronounced in TTC, DAC, and EP, indicating that recursive response reasoning enhances collision avoidance, drivable-area compliance, and driving progress, while Comfort and DDC remain consistently high across all depths. Although Level-5 marginally improves NC and EP, it yields no further gain in PDMS and slightly reduces DAC. These results suggest that deeper reasoning progressively captures strategic dependencies among traffic participants, but exhibits diminishing returns once the dominant interaction structure has been resolved.

Benefits across interaction intensities. To assess whether Level-k reasoning yields greater benefits in interactive settings, we partition all valid tokens into five interaction-intensity groups (None,

Table 2: The performance of the proposed method across different reasoning levels.

Interaction Level	NC	DAC	TTC	Comfort	EP	DDC	PDMS
Level-1	99.06	97.39	96.25	99.97	83.28	98.12	90.03
Level-2	99.25	97.68	97.13	99.97	83.44	**98.17**	90.48
Level-3	99.26	97.74	97.13	**99.98**	83.54	98.15	90.55
Level-4	99.25	**97.81**	**97.18**	99.96	83.60	98.14	**90.64**
Level-5	**99.29**	97.77	**97.18**	99.97	**83.64**	98.16	**90.64**

Table 3: PDMS scores across interaction-intensity groups on NAVTEST. Share denotes the percentage of the 12,144 evaluated tokens, and parentheses report ReWAM's absolute improvement over each baseline.

Interaction Intensity	Tokens	Share (%)	ReWAM	DriveLaW	DiffusionDrive
None	6621	54.52	**89.92**	88.40 (+1.52)	87.06 (+2.86)
Weak	1841	15.16	**90.89**	89.76 (+1.13)	89.20 (+1.69)
Medium	1841	15.16	**92.33**	91.73 (+0.60)	89.56 (+2.77)
Strong	1288	10.61	**92.43**	90.50 (+1.93)	90.17 (+2.26)
Critical	553	4.55	**88.72**	84.13 (+4.59)	86.18 (+2.54)

Table 4: Performance of the proposed method and baselines on critical-interaction scenarios.

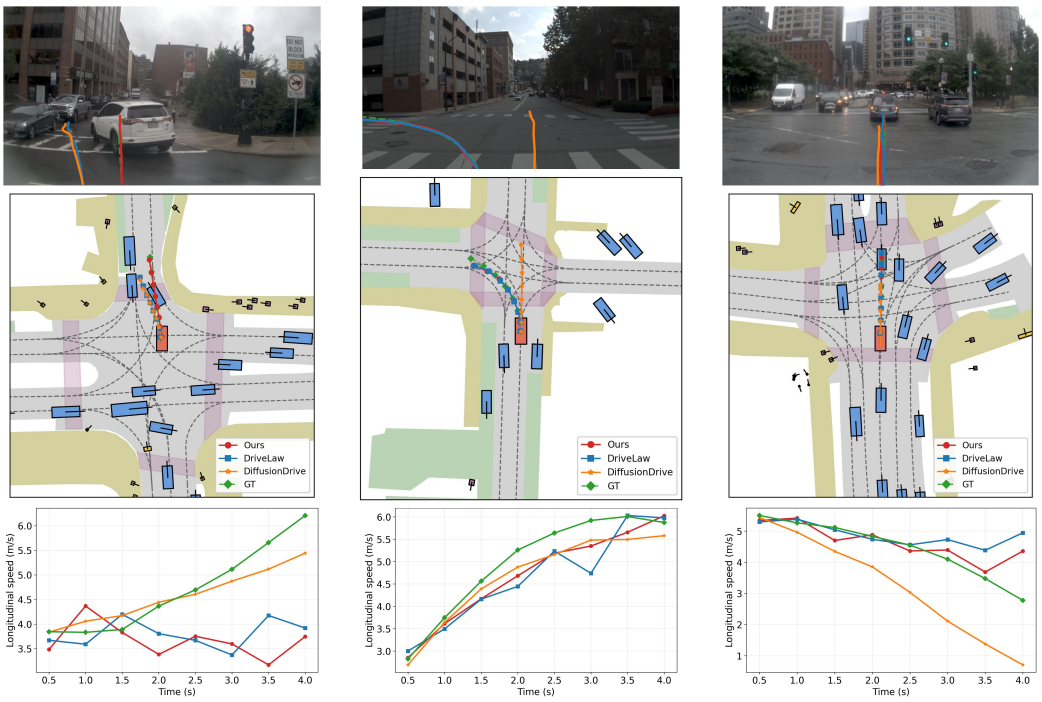
Method	NC	DAC	TTC	Comfort	EP	DDC	PDMS
ReWAM	**96.93**	**99.10**	**85.90**	**100**	**89.59**	**98.19**	**88.72**
DriveLaW	94.21	98.55	79.20	**100**	87.06	98.01	84.13
DiffusionDrive	95.39	97.47	84.27	**100**	87.07	97.38	86.18

Weak, Medium, Strong, and Critical) using a model-independent score derived from the official reference trajectories. The detailed scoring and stratification criteria are provided in Appendix B. As shown in Table 3, ReWAM consistently outperforms both baselines across all five groups, with its advantage becoming particularly pronounced in the most interaction-intensive scenarios. Specifically, its improvement over DriveLaW increases from 1.52 PDMS scores in the None group to 4.59 in the Critical group, where it also surpasses DiffusionDrive by 2.54. Such concentration of performance gains in challenging interactive scenarios provides strong empirical support for the effectiveness of explicitly modeling reciprocal responses among traffic participants.

Performance decomposition under dense interactions. We further examines the Critical subset and the results are shown in Table 4. ReWAM exhibits component-wise dominance over both baselines across all non-saturated metrics, while retaining the same perfect Comfort score. The concurrent improvements in NC and TTC indicate a stronger capacity to anticipate and resolve spatiotemporal motion conflicts. Importantly, these safety gains are accompanied by higher DAC, DDC, and EP, suggesting that conflict avoidance is achieved without resorting to off-road, directionally inconsistent, or excessively conservative behavior. The resulting PDMS improvement therefore reflects a more favorable coordination of safety, traffic-rule compliance, and driving progress, rather than optimization of any single metric. This consistent performance profile provides further evidence that explicit reciprocal-response reasoning is particularly effective when participants' decisions are strongly interdependent.

3.4 Qualitative Results

The results of three interactive scenarios, including front-camera-view action projection, bird's-eye-view action visualization, and longitudinal speed profile, are shown in Figure 4. In the narrow road scenario, ReWAM successfully negotiates the constrained passage, whereas both baselines fail to complete the traversal. In the left turn scenario, DiffusionDrive maintains an almost straight trajectory and misses the intended maneuver, while ReWAM conforms to the turning geometry and preserves smooth progress. The going straight scenario further exposes two opposing failure modes: DiffusionDrive decelerates excessively and sacrifices ego progress, whereas DriveLaW responds insufficiently and collides with the leading vehicle. By contrast, ReWAM adopts a properly calibrated



(a) Narrow Road (b) Left Turn (c) Going Straight

Figure 4: Qualitative comparison across three representative interactive driving scenarios. Each panel presents the front-camera action projection, bird's-eye-view trajectory visualization, and longitudinal speed profile.

trajectory and speed profile that avoids the conflict without unnecessary loss of progress. These examples collectively demonstrate the superior performance of ReWAM in interactive driving scenarios, where reciprocal response modeling yields safer and more effective action.

4 Conclusion

In this paper, we presented the proposed ReWAM, a game-theoretic world-action modeling framework that explicitly captures the reciprocal influence between the action of the ego agent and other agents, advancing WAMs from one-way ego-action conditioning toward strategically coupled multi-agent action generation. The proposed formulation replaces simultaneous multi-agent coupling with a finite hierarchy of bounded responses, thereby introducing structured strategic reasoning without incurring online equilibrium computation. Within this hierarchy, the Ego and parameter-shared Other Action DiTs are grounded in a shared future-world representation and exchange preceding-level action hypotheses through cross-agent attention. Conditional flow matching then learns the corresponding response distributions directly from demonstrations. ReWAM achieves the best performance on NAVSIM, with a PDMS of 90.6, while the study on reasoning depth shows that recursive updates provide consistent gains before saturating at deeper levels. Interaction-stratified evaluation further demonstrates that our method outperforms both DriveLaW and DiffusionDrive across all intensity groups, with its largest advantage over the DriveLaW backbone arising in the critical interaction group. The accompanying improvements in safety, traffic-rule compliance, and ego progress, together with the qualitative examples, validate that these gains arise from more effective interaction-aware action generation rather than conservative planning alone. These findings establish bounded reciprocal-response reasoning as a practical and effective foundation for interactive world-action modeling.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, LaTeX formatting, and code inspection. All technical content, experimental designs, implementations, results, and claims were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We support reproducibility by providing the complete model formulation, architecture, training objective, and inference procedure in Section 2. The observation and trajectory parameterizations, together with the theoretical analysis of conditional flow matching, are detailed in Appendix A.1 and Appendix A.2. Section 3.1 specifies the model initialization, trainable components, input configuration, prediction horizon, dataset splits, and evaluation metrics. The complete interaction-intensity scoring procedure, including eligibility criteria, thresholds, and group distributions, is provided in Appendix B.

REFERENCES

Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, et al. WorldVLA: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025.

Hao Chen, Jiaming Liu, Chenyang Gu, Zhuoyang Liu, Renrui Zhang, Xiaoqi Li, Xiao He, Yandong Guo, Chi-Wing Fu, Shanghang Zhang, et al. Fast-in-Slow: A dual-system vla model unifying fast manipulation within slow reasoning. *Advances in Neural Information Processing Systems*, 38: 98049–98083, 2026a.

Jialei Chen, Kai Wang, Kang Chen, Shuaihang Chen, Feng Gao, Wenhao Tang, Zhiyuan Li, Weilin Liu, Zhuyu Yao, Boxun Li, et al. LaWAM: Latent world action models for efficient dynamics-aware robot policies. *arXiv preprint arXiv:2606.15768*, 2026b.

Xinpeng Chen, Binjie Mou, Zeyu Yang, Yougang Bian, Manjiang Hu, and Zhijun Chen. Contingency planning for autonomous vehicles at uncontrolled intersections based on level-k game model predictions. *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–13, 2026c.

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

Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12878–12895, 2023.

Jingjing Fan, Yushan Liu, Shoujie Li, Botao Ren, Siyuan Li, Xiao-Ping Zhang, Wenbo Ding, and Zhidong Deng. Future-VLA: Forecasting unified trajectories under real-time execution. *arXiv preprint arXiv:2602.15882*, 2026.

Hongwei Fang and Peng Yi. Adaptive decision-making in attack-defense games with bayesian inference of rationality level. *IEEE Transactions on Industrial Electronics*, 71(12):16558–16567, 2024.

Xingtai Gui, Meijie Zhang, Tianyi Yan, Wencheng Han, Jiahao Gong, Feiyang Tan, Cheng zhong Xu, and Jianbing Shen. Bridging scene generation and planning: Driving with world model via unifying vision and motion representation. *arXiv preprint arXiv:2603.14948*, 2026.

Ke Guo, Haochen Liu, Xiaojun Wu, Jia Pan, and Chen Lv. iPad: Iterative proposal-centric end-to-end autonomous driving. *IEEE Robotics and Automation Letters*, 2026.

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

Zhiyu Huang, Haochen Liu, and Chen Lv. GameFormer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving. In *2023 IEEE/CVF International Conference on Computer Vision*, pp. 3880–3890, 2023.

Feiyang Jia, Lin Liu, Ziying Song, Caiyan Jia, Hangjun Ye, Xiaoshuai Hao, and Long Chen. Driveworld-VLA: Unified latent-space world modeling with vision–language–action for autonomous driving. In *Forty-third International Conference on Machine Learning*, 2026.

Peng Jia, Jianwei Gong, Lanheng Nie, Zhiyang Ju, Ruizeng Zhang, Lei Tian, and Tao Qin. Ipp: Interactive policy planning with adaptive trajectory optimization and joint conditional prediction. *IEEE Transactions on Consumer Electronics*, 2025a.

Xiaosng Jia, Junqi You, Zhiyuan Zhang, and Junchi Yan. DriveTransformer: Unified transformer for scalable end-to-end autonomous driving. pp. 57196 – 57212, Singapore, Singapore, 2025b.

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

Sicong Jiang, Zilin Huang, Kangan Qian, Ziang Luo, Tianze Zhu, Yang Zhong, Yihong Tang, Menglin Kong, Yunlong Wang, Siwen Jiao, et al. A survey on vision-language-action models for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4583–4595, 2025.

Shahab Karimi, Arash Karimi, and Ardalan Vahidi. Level-k reasoning, deep reinforcement learning, and monte carlo decision process for fast and safe automated lane change and speed management. *IEEE Transactions on Intelligent Vehicles*, 8(6):3556–3571, 2023.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. volume 270, pp. 2679 – 2713, 2024.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Xunhao Li, Jian Zhang, Jun Chen, and Pinzheng Qian. Interaction-aware trajectory prediction method based on sparse spatial-temporal transformer for internet of vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 26(8):12025–12038, 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. pp. 34849 – 34866, 2025b.

Yingyan Li, Yuqi Wang, Yang Liu, Jiawei He, Lue Fan, and Zhaoxiang Zhang. End-to-end driving with online trajectory evaluation via bev world model. In *2025 IEEE/CVF International Conference on Computer Vision*, pp. 27137–27146, 2025c.

Yingyan Li, Shuyao Shang, Weisong Liu, Bing Zhan, Haochen Wang, Yuqi Wang, Yuntao Chen, Xiaoman Wang, AnYasong, Chufeng Tang, Lu Hou, Lue Fan, and Zhaoxiang Zhang. DriveVLA-w0: World models amplify data scaling law in autonomous driving. In *The Fourteenth International Conference on Learning Representations*, 2026.

Jinhao Liang, Chaopeng Tan, Longhao Yan, Jingyuan Zhou, Guodong Yin, and Kaidi Yang. Interaction-aware trajectory prediction for safe motion planning in autonomous driving: A transformer-transfer learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 2025.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12037–12047, 2025.

Pei Liu, Haipeng Liu, Haichao Liu, Xin Liu, Jinxin Ni, and Jun Ma. Vlm-e2e: Enhancing end-to-end autonomous driving with multimodal driver attention fusion. *arXiv preprint arXiv:2502.18042*, 2025.

Pei Liu, Qingtian Ning, Xinyan Lu, Haipeng Liu, Weiliang Ma, Dangen She, Xianpeng Lang, and Jun Ma. CogDriver: Integrating cognitive inertia for temporally coherent planning in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18150–18160, 2026.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworld-model: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

Baorui Peng, Wenyao Zhang, Liang Xu, Zekun Qi, Jiazhao Zhang, Hongsi Liu, Wenjun Zeng, and Xin Jin. Reworld: Multi-dimensional reward modeling for embodied world models. *arXiv preprint arXiv:2601.12428*, 2026.

Shuyao Shang, Yuntao Chen, Yuqi Wang, Yingyan Li, and Zhaoxiang Zhang. DriveDPO: Policy learning via safety DPO for end-to-end autonomous driving. *Advances in Neural Information Processing Systems*, 38:81565–81585, 2026.

Jingwen Sun, Wenyao Zhang, Zekun Qi, Shaojie Ren, Zezhi Liu, Hanxin Zhu, Guangzhong Sun, Xin Jin, and Zhibo Chen. VLA-JEPA: Enhancing vision-language-action model with latent world model. *arXiv preprint arXiv:2602.10098*, 2026.

Tianyi Tan, Yinan Zheng, Ruiming Liang, Zexu Wang, Kexin Zheng, Jinliang Zheng, Jianxiong Li, Xianyuan Zhan, and Jingjing Liu. Flow matching-based autonomous driving planning with advanced interactive behavior modeling. *Advances in Neural Information Processing Systems*, 38:38310–38335, 2026.

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

Junli Wang, Yinan Zheng, Xueyi Liu, Zebin Xing, Pengfei Li, Kun Ma, Hangjun Ye, Guang Chen, Guang Li, Long Chen, et al. MeanFuser: Fast one-step multi-modal trajectory generation and adaptive reconstruction via meanflow for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17884–17893, 2026b.

Linbo Wang, Yupeng Zheng, Qiang Chen, Shiwei Li, Yichen Zhang, Zebin Xing, Qichao Zhang, Xiang Li, Deheng Qian, Pengxuan Yang, et al. Latent-WAM: Latent world action modeling for end-to-end autonomous driving. *arXiv preprint arXiv:2603.24581*, 2026c.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22442–22452, 2025.

Xinshuo Weng, Boris Ivanovic, Yan Wang, Yue Wang, and Marco Pavone. PARA-Drive: Parallelized architecture for real-time autonomous driving. In *2024 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, et al. 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.

Wenli Xiao, Haotian Lin, Andy Peng, Haoru Xue, Tairan He, Zhengyi Luo, Yuqi Xie, Fengyuan Hu, Jim Fan, Guanya Shi, et al. Self-improving vision-language-action models with data generation via residual rl. In *International Conference on Learning Representations*, volume 2026, pp. 13209–13236, 2026.

Chengen Xie, Chonghao Sima, Tianyu Li, Bin Sun, Junjie Wu, Zhihui Hao, and Hongyang Li. Flare: Learning future-aware latent representations from vision-language models for autonomous driving. *arXiv preprint arXiv:2601.05611*, 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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1602–1611, 2025.

Kaixin Xiong, Xiangyu Guo, Fang Li, Sixu Yan, Gangwei Xu, Lijun Zhou, Long Chen, Haiyang Sun, Bing Wang, Kun Ma, et al. ReCogDrive: A reinforced cognitive framework for end-to-end autonomous driving. In *International Conference on Learning Representations*, volume 2026, pp. 157518–157556, 2026.

Fuxiang Yang, Donglin Di, Lulu Tang, Xuancheng Zhang, Lei Fan, Hao Li, Wei Chen, Tonghua Su, and Baorui Ma. Chain of World: World model thinking in latent motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6675–6684, 2026.

Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Hao Li, Hengtao Li, Jie Li, Jindi Lv, Jingyu Liu, et al. GigaWorld-Policy: An efficient action-centered world–action model. *arXiv preprint arXiv:2603.17240*, 2026a.

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

Mingfeng Yuan, Jinjun Shan, and Hunter Schofield. Scalable game-theoretic decision-making for self-driving cars at unsignalized intersections. *IEEE Transactions on Industrial Electronics*, 71 (6):5920–5930, 2023.

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

Dapeng Zhang, Zhenlong Yuan, Zhangquan Chen, Chih-Ting Liao, Yinda Chen, Fei Shen, Qingguo Zhou, and Tat-Seng Chua. Reasoning-VLA: An efficient and spatial-guided general vision-language-action reasoning model for autonomous driving. In *Forty-third International Conference on Machine Learning*, 2026.

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*, 2025.

Zhida Zhao, Talas Fu, Yifan Wang, Lijun Wang, and Huchuan Lu. From forecasting to planning: Policy world model for collaborative state-action prediction. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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 *2025 IEEE/CVF International Conference on Computer Vision*, pp. 28632–28642, 2025.

Xingcheng Zhou, Xuyuan Han, Feng Yang, Yunpu Ma, Volker Tresp, and Alois Knoll. OpenDriveVLA: Towards end-to-end autonomous driving with large vision language action model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13782–13790, 2026a.

Yang Zhou, Xiaofeng Wang, Hao Shao, Letian Wang, Guosheng Zhao, Jiangnan Shao, Jiagang Zhu, Tingdong Yu, Zheng Zhu, Guan Huang, et al. DriveDreamer-Policy: A geometry-grounded world-action model for unified generation and planning. *arXiv preprint arXiv:2604.01765*, 2026b.

Zewei Zhou, Tianhui Cai, Seth Zhao, Yun Zhang, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. AutoVLA: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. *Advances in Neural Information Processing Systems*, 38: 27920–27956, 2026c.

A Additional Information

A.1 Observation and Trajectory Parameterizations

We omit the batch dimension throughout for clarity. At time t, the model observation is

$$\mathcal{O}_t = \left\{ I_{t-H:t}, c_t, s_t^e, \{s_t^i\}_{i=1}^N \right\},$$

where $I_{t-H:t} \in \mathbb{R}^{(H+1)\times C\times H_I\times W_I}$ denotes the historical visual observations, $c_t \in \mathbb{R}^{L_c\times d_c}$ is the encoded navigation or language condition, and $s_t^e \in \mathbb{R}^{d_e}$ is the ego state. Each surrounding state is represented as

$$s_t^i = [\Delta x, \Delta y, \sin\theta, \cos\theta, v_x, v_y, l, w] \in \mathbb{R}^8,$$

which contains its relative pose, velocity, and bounding-box size.

The ego decoder produces $a_e = [x, y, \theta]_{1:T} \in \mathbb{R}^{T\times 3}$. For interaction reasoning, we transform it into the canonical strategy representation

$$\tau_e = g(a_e) = [x, y, \sin\theta, \cos\theta]_{1:T} \in \mathbb{R}^{T\times 4}. \tag{19}$$

The surrounding branch directly represents the strategy of each surrounding agent as $\tau_i \in \mathbb{R}^{T\times 4}$. Consequently, ego and surrounding strategies share the same representation when exchanged across reasoning levels.

A.2 Consistency of Joint Conditional Flow Matching

After excluding invalid agents, the aggregate objective in equation 16 is a positively weighted combination of the participant-level population risks:

$$\mathcal{L}_{\text{flow}}(\Theta) = \sum_{(j,k)\in\mathcal{I}} \beta_{j,k}\mathcal{L}_j^{(k)}(\Theta), \qquad \beta_{j,k} > 0, \tag{20}$$

where $\mathcal{I}$ indexes all valid participants and supervised reasoning levels. The coefficients $\beta_{j,k}$ incorporate the normalized reasoning weight α_k, the corresponding branch weight, and the validity normalization in equation 15 and equation 16.

We assume simultaneous realizability: a shared parameterization $\bar{\Theta}$ attains the population infimum of every component risk,

$$\mathcal{L}_j^{(k)}(\bar{\Theta}) = \inf_{\Theta} \mathcal{L}_j^{(k)}(\Theta), \qquad (j,k) \in \mathcal{I}. \tag{21}$$

Because every $\beta_{j,k}$ is positive, any population minimizer $\Theta^\star$ of the aggregate objective must also minimize every component risk. Otherwise, its aggregate risk would be strictly larger than the value attained by $\bar{\Theta}$. This establishes the connection between the joint optimization in equation 17 and the participant-level objective in equation 14.

Fix a valid participant j and supervised reasoning level k, and abbreviate $(\mathbf{W}_t, s_t^j, \mathbf{Z}_{-j}^{(k-1)})$ by $\mathbf{c}$. Let $p_0(\cdot \mid \mathbf{c})$ be the normalized demonstration distribution and let $p_1 = \mathcal{N}(\mathbf{0}, \mathbf{I})$. Draw $\mathbf{x}_0 \sim p_0(\cdot \mid \mathbf{c})$ and $\mathbf{x}_1 \sim p_1$, and define $\mathbf{x}_r = (1-r)\mathbf{x}_0 + r\mathbf{x}_1$ and $\mathbf{u}_r = \mathbf{x}_1 - \mathbf{x}_0$.

The ℓ_2 projection property of conditional regression gives

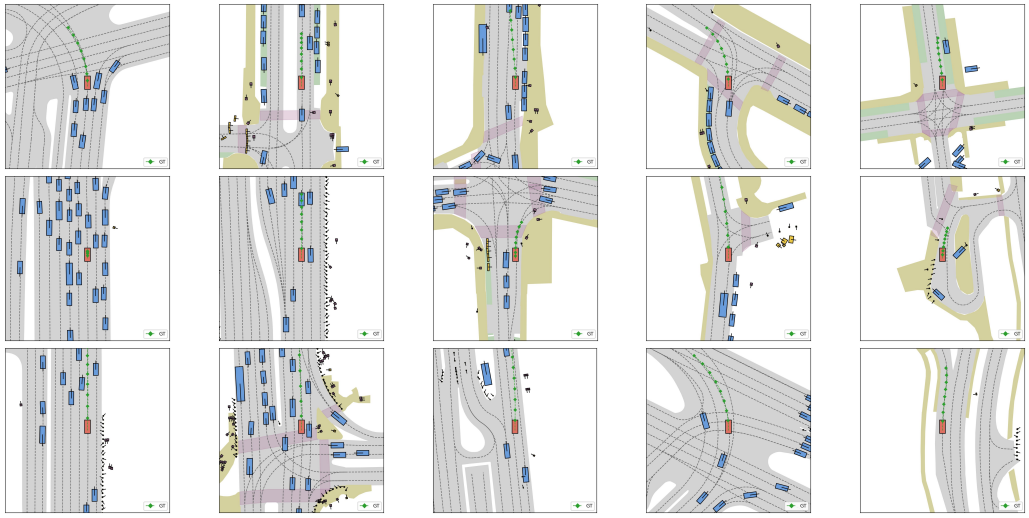
$$v^\star(\mathbf{x}, r \mid \mathbf{c}) = \mathbb{E}[\mathbf{u}_r \mid \mathbf{x}_r = \mathbf{x}, \mathbf{c}]. \tag{22}$$

For any smooth compactly supported test function φ,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}r}\mathbb{E}[\varphi(\mathbf{x}_r) \mid \mathbf{c}] &= \mathbb{E}\big[\nabla\varphi(\mathbf{x}_r)^\top \mathbf{u}_r \mid \mathbf{c}\big] \\ &= \int \nabla\varphi(\mathbf{x})^\top v^\star(\mathbf{x}, r \mid \mathbf{c}) p_r(\mathbf{x} \mid \mathbf{c})\,\mathrm{d}\mathbf{x}. \end{aligned} \tag{23}$$

Integration by parts shows that p_r satisfies

$$\partial_r p_r(\mathbf{x} \mid \mathbf{c}) + \nabla_{\mathbf{x}}\cdot(p_r(\mathbf{x} \mid \mathbf{c}) v^\star(\mathbf{x}, r \mid \mathbf{c})) = 0. \tag{24}$$



(a) Critical (b) Strong (c) Medium (d) Weak (e) None

Figure 5: NAVTEST scenario examples across interaction-intensity groups, ordered from Critical to None.

This continuity equation has endpoints $p_0(\cdot \mid \mathbf{c})$ and p_1. Therefore, integrating $\mathrm{d}\mathbf{x}_r/\mathrm{d}r = v^\star(\mathbf{x}_r, r \mid \mathbf{c})$ from $r = 1$ to $r = 0$ transports the Gaussian prior to the conditional demonstration distribution. Under the regularity and realizability assumptions above, the branch parameters $\theta_j^\star$ inherited from $\Theta^\star$ therefore satisfy

$$q_{\theta_j^\star}(\cdot \mid \mathbf{c}) = p_0(\cdot \mid \mathbf{c}) \quad \text{for all } (j,k) \in \mathcal{I}, \tag{25}$$

almost surely over the training-context distribution. This proves equation 18 jointly across valid participants and supervised reasoning levels. Figure 5 provides representative examples from the five interaction-intensity groups.

B Interaction-Intensity Stratification

B.1 Trajectory-Based Interaction Measures

Interaction labels are computed exclusively from the official NAVTEST metric cache and are therefore independent of the predictions and scores of all compared methods. For each token, we use the reference ego trajectory and the trajectories of surrounding objects labeled as vehicles over a $4.0\,\mathrm{s}$ horizon with sampling interval $\Delta t = 0.5\,\mathrm{s}$. Let P_m^e and P_m^a denote the polygons of the ego vehicle and surrounding vehicle a at time t_m, respectively. All distances below are minimum polygon-to-polygon distances rather than distances between vehicle centers. Time indices at which vehicle a is absent are excluded from the corresponding computation; if it is absent at the current time, we set $d_{\mathrm{cur}}^a = +\infty$.

For each surrounding vehicle, we compute its current distance, displacement, and peak speed as

$$d_{\mathrm{cur}}^a = \mathrm{dist}(P_0^e, P_0^a), \tag{26}$$

$$d_{\mathrm{disp}}^a = \max_m \|\mathbf{c}_m^a - \mathbf{c}_0^a\|_2, \tag{27}$$

$$v_{\mathrm{peak}}^a = \max_{m \geq 1} \frac{\|\mathbf{c}_m^a - \mathbf{c}_{m-1}^a\|_2}{\Delta t}, \tag{28}$$

where $\mathbf{c}_m^a$ is the center of P_m^a. We further define the synchronous minimum distance and maximum closing speed by

$$d_{\mathrm{sync}}^a = \min_m \mathrm{dist}(P_m^e, P_m^a), \tag{29}$$

$$v_{\mathrm{closing}}^a = \max_{m \geq 1} \left[-\frac{d_m^a - d_{m-1}^a}{\Delta t} \right]_+, \qquad d_m^a = \mathrm{dist}(P_m^e, P_m^a), \tag{30}$$

Table 5: Definition and distribution of NAVTEST interaction-intensity groups. Share is measured relative to the 12,144 jointly valid tokens.

Group	Criterion	Tokens	Share (%)
None	$N_{\text{int}} = 0$	6,621	54.52
Weak	$0 < I_{\text{scene}} \leq q_{1/3}$	1,841	15.16
Medium	$q_{1/3} < I_{\text{scene}} \leq q_{2/3}$	1,841	15.16
Strong	$q_{2/3} < I_{\text{scene}} \leq q_{0.9}$	1,288	10.61
Critical	$I_{\text{scene}} > q_{0.9}$	553	4.55

with $[x]_+ = \max(x, 0)$. To capture conflicts in which two vehicles reach the same spatial region at different times, we additionally compute

$$d^a_{\text{path}} = \min_{m,n} \text{dist}(P^e_m, P^a_n). \tag{31}$$

When $d^a_{\text{path}} \leq 3.0\,\text{m}$, define $d^a_{\text{region}} = \min(3.0\,\text{m}, d^a_{\text{path}} + 0.25\,\text{m})$ and

$$\Delta t^a_{\text{conflict}} = \min_{\substack{m,n: \\ \text{dist}(P^e_m, P^a_n) \leq d^a_{\text{region}}}} |t_m - t_n|. \tag{32}$$

Otherwise, $\Delta t^a_{\text{conflict}} = +\infty$.

A surrounding vehicle is eligible for interaction scoring if it is moving, currently nearby, and either becomes synchronously close to the ego vehicle or exhibits a spatiotemporal path conflict. Formally,

$$\text{moving}^a \equiv (d^a_{\text{disp}} \geq 1.0\,\text{m}) \vee (v^a_{\text{peak}} \geq 0.5\,\text{m/s}), \tag{33}$$

$$\text{eligible}^a \equiv \text{moving}^a \wedge (d^a_{\text{cur}} \leq 35.0\,\text{m}) \wedge \Big[(d^a_{\text{sync}} \leq 3.0\,\text{m}) \vee \big(d^a_{\text{path}} \leq 3.0\,\text{m} \wedge \Delta t^a_{\text{conflict}} \leq 1.5\,\text{s}\big)\Big]. \tag{34}$$

B.2 Scene Score and Intensity Groups

For every eligible vehicle, four normalized terms quantify synchronous proximity, spatiotemporal path conflict, closing speed, and vehicle motion:

$$S^a_{\text{prox}} = \exp\left(-\frac{d^a_{\text{sync}}}{3.0\,\text{m}}\right), \tag{35}$$

$$S^a_{\text{st}} = \exp\left(-\frac{d^a_{\text{path}}}{3.0\,\text{m}}\right) \exp\left(-\frac{\Delta t^a_{\text{conflict}}}{1.5\,\text{s}}\right), \tag{36}$$

$$S^a_{\text{closing}} = \text{clip}\left(\frac{v^a_{\text{closing}}}{5.0\,\text{m/s}}, 0, 1\right), \tag{37}$$

$$S^a_{\text{motion}} = \text{clip}\left(\frac{v^a_{\text{peak}}}{8.0\,\text{m/s}}, 0, 1\right). \tag{38}$$

Here, $\text{clip}(x, 0, 1) = \min(\max(x, 0), 1)$. The per-vehicle interaction intensity is

$$I_a = 0.35 S^a_{\text{prox}} + 0.35 S^a_{\text{st}} + 0.20 S^a_{\text{closing}} + 0.10 S^a_{\text{motion}}. \tag{39}$$

If N_{int} vehicles satisfy the eligibility criterion, the scene score is defined as

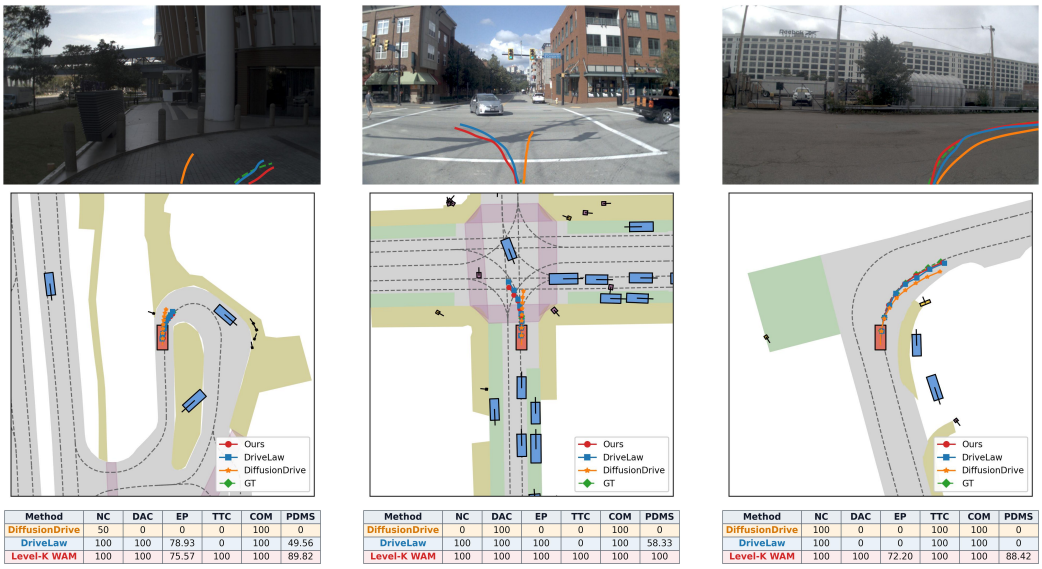
$$I_{\text{scene}} = \text{clip}\left(\max_{a:\,\text{eligible}^a} I_a + 0.05 \min\big(\max(N_{\text{int}} - 1, 0), 2\big), 0, 1\right). \tag{40}$$

The second term assigns bonuses of 0.05 and 0.10 to scenes with two and at least three eligible vehicles, respectively. We set $I_{\text{scene}} = 0$ when $N_{\text{int}} = 0$.

Among the 12,144 tokens jointly valid for all evaluated methods, 6,621 contain no eligible interacting vehicle and are assigned to the *None* group. The remaining 5,523 positive scores are partitioned at their one-third, two-thirds, and 90th percentiles,

$$q_{1/3} = 0.5170426573, \qquad q_{2/3} = 0.7379162662, \qquad q_{0.9} = 0.9146632565. \tag{41}$$

The resulting stratification is summarized in Table 5.



Method	NC	DAC	EP	TTC	COM	PDMS
DiffusionDrive	50	0	0	0	100	0
DriveLaw	100	100	78.93	0	100	49.56
Level-K WAM	100	100	75.57	100	100	89.82

(a) Case 1

Method	NC	DAC	EP	TTC	COM	PDMS
DiffusionDrive	0	100	0	0	100	0
DriveLaw	100	100	100	0	100	58.33
Level-K WAM	100	100	100	100	100	100

(b) Case 2

Method	NC	DAC	EP	TTC	COM	PDMS
DiffusionDrive	100	0	0	100	100	0
DriveLaw	100	0	0	100	100	0
Level-K WAM	100	100	72.20	100	100	88.42

(c) Case 3

Figure 6: Additional qualitative comparisons in three representative interactive driving scenarios. Each panel presents the front-camera action projection, bird's-eye-view trajectory visualization, and the corresponding planning metrics.

C More Cases

As shown in Figure 6, the three cases reveal complementary failure modes that are not apparent from trajectory geometry alone. In the first case, ReWAM achieves a PDMS of 89.82 by attaining full NC, DAC, TTC, and Comfort scores while preserving an EP of 75.57. DriveLaW achieves slightly higher progress (78.93) but receives a TTC score of zero, reducing its PDMS to 49.56, whereas DiffusionDrive obtains zero PDMS because of simultaneous failures in DAC, TTC, and EP. In the second case, ReWAM is the only method to achieve perfect scores across all components. Although DriveLaW preserves collision avoidance, route compliance, and progress, its zero TTC score limits PDMS to 58.33; DiffusionDrive fails both collision- and interaction-related metrics. In the third case, both baselines obtain zero EP, and DiffusionDrive additionally violates drivable-area compliance. By contrast, ReWAM maintains perfect NC, DAC, TTC, and Comfort scores while attaining 72.20 EP and 88.42 PDMS. Collectively, these results show that ReWAM more reliably resolves interaction-induced conflicts and generates feasible trajectories that sustain meaningful driving progress.