ReVAD: From Imitation to Reasoning in Vectorized Autonomous Driving via Latent Space Search

Zhao Yang[*1] **Chengkang Duan**[*1] **Weiyi Hu**[2] **Haoran Hu**[1] **Hua Cui**[†1] **Qingshuang Sun**[3]

[1]School of Future Transportation, Chang'an University, Xi'an, Shaan'xi, China
[2]School of Information Engineering, Chang'an University, Xi'an, Shaan'xi, China
[3]College of Computer Science, Xi'an Polytechnic University, Xi'an, Shaan'xi, China

Abstract

Vectorized end-to-end (E2E) autonomous driving models predominantly rely on reactive imitation learning. By failing to model underlying environmental dynamics, these agents lack counterfactual reasoning capabilities and struggle with uncertainty in out-of-distribution (OOD) scenarios. To bridge the gap between reactive imitation and deliberative decision-making, we introduce Reasoning-enhanced VAD (ReVAD), empowering agents with a "System 2" cognitive framework via Latent Space Search. Unlike computationally expensive pixel-level world models, ReVAD integrates a probabilistic Token Dynamics Model with Latent Monte Carlo Tree Search to perform efficient lookahead planning entirely within a sparse semantic token space. By simulating future states to evaluate risk and utilizing imitation policies solely as search priors, ReVAD effectively mitigates the distribution shift inherent in pure imitation learning, demonstrating significantly improved robustness and safety in high-uncertainty environments.

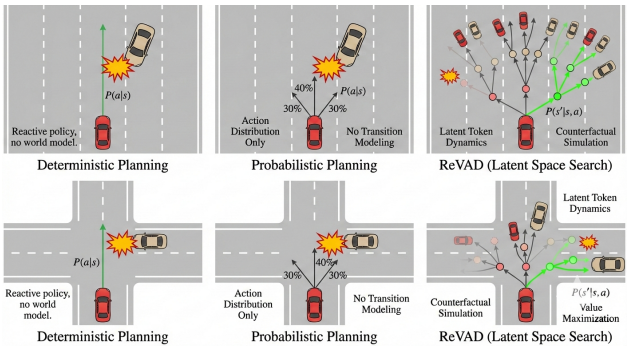

Figure 1: Deterministic and probabilistic planning both belong to imitation-based System 1 paradigms, differing only in how they represent the action distribution $P(a \mid s)$, while sharing the absence of an explicit transition model. Without modeling the environment dynamics $P(s' \mid s, a)$, they cannot perform counterfactual reasoning under uncertainty. In contrast, ReVAD integrates probabilistic token dynamics with latent Monte Carlo Tree Search (MCTS), enabling deliberative planning through simulated future state evolution.

1 INTRODUCTION

The paradigm of autonomous driving is currently navigating a critical transition from modular pipelines to End-to-End (E2E) learning. Within this domain, vectorized approaches like VADv2 Jiang et al. [2026] have established a new state-of-the-art (SOTA) by representing the driving scene as sparse, instance-level tokens rather than dense rasterized images. While this tokenization offers significant advantages in inference speed and structural interpretability, the underlying decision-making mechanism of these models remains fundamentally rooted in Imitation Learning. Operationalizing as reflexive "System 1" agents, they approximate the conditional distribution of expert actions given the context, $P(a|s)$, effectively maximizing likelihood on a fixed training manifold without modeling the causal dynamics of the environment. As illustrated in Fig. 1, both deterministic and probabilistic planning in current vectorized E2E systems remain imitation-based System 1 paradigms: they differ mainly in how they parameterize $P(a \mid s)$, yet share the absence of an explicit transition model $P(s' \mid s, a)$.

This reliance on static behavioral priors presents a severe limitation, particularly when handling uncertainty in safety-critical and long-tail scenarios. When varying from the expert's average behavior or encountering out-of-distribution (OOD) events—such as sudden occlusions or aggressive interactions—pure imitation models lack the capacity for counterfactual reasoning. They cannot explicitly query,

* These authors contributed equally to this work.
† Corresponding author. Email: huacui@chd.edu.cn

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:7751–7760.

"*If I execute action a', how will the transition dynamics $P(s'|s, a')$ unfold?*" Consequently, without an internal world model to simulate future states and evaluate risk, these agents are prone to making confident yet catastrophic errors when the observation drifts from the training distribution.

To bridge the gap between reactive imitation and deliberative reasoning, we propose ReVAD (Reasoning-enhanced VAD), a vectorized driving framework equipped with Latent Space Search. Unlike pixel-space generative world models, which are often too costly for real-time planning and may suffer from visual hallucinations, ReVAD performs lookahead planning entirely in a learned Latent Token Space. This choice follows recent progress on compact latent predictive models for driving, including latent forecasting for E2E planning Li et al. [2025], multiview world-model forecasting Wang et al. [2024b], and world-model-based representation learning Min et al. [2024], as well as intention-aware latent world models for multimodal planning Zheng et al. [2025]. We argue that instance-level tokens provide an effective abstraction for planning: semantically expressive for multi-agent interaction, yet sparse enough for efficient, high-frequency rollouts.

The core of ReVAD is the integration of a probabilistic Token Dynamics Model with Latent MCTS. We design a Transformer-based dynamics model that predicts the evolution of heterogeneous tokens (map elements and traffic agents) conditioned on ego-actions. During inference, ReVAD utilizes the imitation policy as a prior but validates and refines decisions through online search. This architecture effectively instantiates a "System 2" reasoning process, allowing the agent to "think before acting" by exploring future trajectories and maximizing a comprehensive value function that accounts for safety, compliance, and progress. Our main contributions are summarized as follows:

- We introduce an E2E driving framework that integrates Latent Space Search into a fully vectorized architecture, enabling robust online planning without the computational burden of pixel-level generation.
- We develop a dynamics model capable of predicting future token states, enabling the agent to perform counterfactual analysis and collision checking in the latent space.
- We demonstrate that supplementing imitation priors with model-based search significantly improves robustness in high-uncertainty scenarios, effectively mitigating the "distribution shift" problem inherent in pure imitation learning.

2 RELATED WORK

E2E Driving and the Imitation Bottleneck. The transition from modular pipelines to E2E autonomous driving has been significantly accelerated by vectorized representations. Recent SOTA frameworks, such as VADv2 Jiang et al. [2026], utilize sparse, instance-level tokens to represent complex driving scenes, offering substantial improvements in inference speed and structural interpretability. However, the decision-making core of these models predominantly relies on imitation learning or Behavioral Cloning. Operating as reactive "System 1" agents, they directly approximate the expert's conditional action distribution $P(a \mid s)$. While effective within the training manifold, pure imitation learning methods are notoriously susceptible to compounding errors and covariate shift when encountering OOD scenarios Ross et al. [2011], Codevilla et al. [2019]. Unlike our proposed ReVAD, these approaches lack explicit transition models and are fundamentally incapable of counterfactual reasoning under high-uncertainty conditions.

World Models and Latent Dynamics. To address the lack of environmental understanding in model-free agents, predictive world models have gained traction. Early approaches in autonomous driving focused on high-dimensional, pixel-level video generation to simulate future states Hu et al. [2023a], Wayve [2023]. While these generative models provide rich visual foresight, they are computationally intractable for high-frequency online planning and are highly prone to visual hallucinations that can compromise safety-critical decisions. Recent advancements in reinforcement learning have demonstrated the efficacy of planning within abstract latent spaces Hafner et al. [2020], Schrittwieser et al. [2020], a paradigm shift that is increasingly being adopted in the latest E2E driving architectures to model physical multi-agent interactions without the burden of pixel rendering Zheng et al. [2025], Sun et al. [2025]. Building upon this philosophy, ReVAD avoids pixel-level reconstruction entirely. Instead, it introduces a probabilistic Token Dynamics Model that explicitly learns the transition function $P(s' \mid s, a)$ directly within a sparse semantic token space.

Deliberative Planning and System 2 Reasoning. The integration of planning algorithms with learned neural priors has led to breakthrough performances in complex decision-making domains, famously instantiated by MCTS in AlphaZero and MuZero Silver et al. [2017], Schrittwieser et al. [2020]. In the context of autonomous driving, purely deterministic planning often fails to account for the stochastic nature of human drivers. Consequently, cutting-edge research in 2025 has surged towards dual-system cognitive architectures Hamdan et al. [2025] and the integration of tree-search algorithms for safe urban navigation Momchil et al. [2025]. ReVAD comprehensively extends this paradigm by embedding Latent MCTS into a fully vectorized driving architecture. By utilizing the imitation learning policy as an action prior rather than a definitive oracle, ReVAD dynamically explores simulated future trajectories, effectively

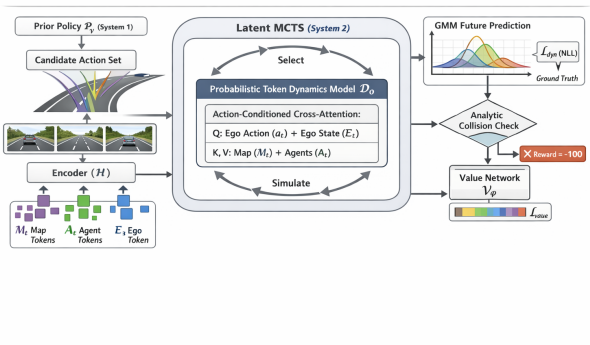

Figure 2: Overall architecture of ReVAD. ReVAD streams multi-view images, tokenizes them into map/agent/ego tokens, and proposes candidate actions via a prior policy. A latent MCTS planner simulates and scores actions with a probabilistic dynamics model, collision check, and value network, then outputs an action distribution and samples one for control. Demonstrations and scene constraints supervise training.

formalizing a deliberative "System 2" cognitive process that maximizes safety and utility through online counterfactual risk assessment.

3 REVAD

To bridge the gap between reactive imitation learning (System 1) and deliberative reasoning (System 2) in E2E autonomous driving, we propose the ReVAD framework, as shown in Fig. 2. In this section, we detail the underlying mathematical formulations and network architectures of ReVAD. We first formalize the driving decision-making process as a Partially Observable Markov Decision Process (POMDP) in a latent space, followed by the definition of the vectorized state representation. Building upon this, we detail the Probabilistic Token Dynamics Model, which serves as the core for counterfactual reasoning, alongside explicit reward and value estimation networks. To stably optimize this complex dual-system architecture, we introduce a decoupled two-stage offline training paradigm. Finally, we derive the online planning mechanism using Latent MCTS during the inference phase.

3.1 PROBLEM FORMULATION & VECTORIZED STATE ABSTRACTION

We formalize vision-based E2E autonomous driving as a discrete-time POMDP, defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \Omega, \mathcal{O}, \gamma)$. At each time step t, the agent receives a high-dimensional observation $o_t \in \Omega$ (e.g., multi-view camera images) and infers the current latent state of the system $s_t \in \mathcal{S}$ based on the observation history.

Unlike previous generative world models where $\mathcal{S}$ is typically defined as a high-dimensional pixel space or opaque latent vectors, ReVAD utilizes the pre-trained VADv2 encoder $\mathcal{H}$ to map the observation sequence into a highly structured, sparse, and semantically hierarchical set of instance-level tokens. Specifically, the latent state s_t comprises three heterogeneous token sets:

$$s_t = \{\mathcal{M}_t, \mathcal{A}_t, \mathcal{E}_t\} \tag{1}$$

1. **Static Map Tokens ($\mathcal{M}_t$):** Represents the road topology (e.g., lane boundaries, crosswalks, stop lines). $\mathcal{M}_t = \{m_i\}_{i=1}^{N_m}$, where each map element $m_i \in \mathbb{R}^{d_m}$ is encoded as a feature vector containing a sequence of 2D coordinates and semantic category labels.
2. **Dynamic Agent Tokens ($\mathcal{A}_t$):** Represents traffic participants (e.g., surrounding vehicles, pedestrians, cyclists). $\mathcal{A}_t = \{a_j\}_{j=1}^{N_a}$. To support rigorous collision detection in subsequent steps, each dynamic token is explicitly decoupled into geometric and kinematic attributes: $a_j = [x_j, y_j, w_j, l_j, \theta_j, v_{x,j}, v_{y,j}, c_j]$, corresponding to center coordinates, width, length, heading angle, lateral/longitudinal velocities, and class.
3. **Ego Kinematic Token ($\mathcal{E}_t$):** Encodes the ego-vehicle's current state and historical trajectory features, $\mathcal{E}_t \in \mathbb{R}^{d_e}$.

The agent's action space $\mathcal{A}$ is continuous. At time step t, the ego-action $a_t \in \mathcal{A}$ is parameterized as a sequence of local trajectory waypoints for the future H_a steps: $a_t = \{(x_{t+k}, y_{t+k})\}_{k=1}^{H_a}$. This instance-level representation perfectly balances computational efficiency with geometric interpretability, laying the foundation for dynamics evolution in the feature space.

3.2 PROBABILISTIC TOKEN DYNAMICS MODEL

To enable System 2 lookahead planning, the agent must comprehend environmental causality. We introduce the Probabilistic Token Dynamics Model, $\mathcal{D}_\theta$, which acts as the agent's internal world simulator. Given the current state s_t and a candidate action a_t, $\mathcal{D}_\theta$ predicts the environmental state transition probability for the next discrete time step:

$$P(s_{t+1}|s_t, a_t) = \mathcal{D}_\theta(s_t, a_t) \tag{2}$$

Given the unordered nature of token sets and the complexity of multi-agent interactions, we employ a Transformer-based architecture to instantiate $\mathcal{D}_\theta$. We decompose this transition into two parts: the constancy of the static map (translation/rotation mapping in the ego-coordinate system) and the interactive evolution of dynamic agents.

Cross-Attention and Interaction Modeling: To predict how surrounding vehicles will react to the ego-vehicle's actions (e.g., a rear vehicle decelerating if the ego-vehicle forcefully changes lanes), we design an Action-conditioned Cross-Attention mechanism. The fused representation of the ego-action a_t and the ego-state $\mathcal{E}_t$ serves as the Query (Q), while the environmental agent tokens $\mathcal{A}_t$ and map tokens $\mathcal{M}_t$ serve as Keys (K) and Values (V):

$$Q_t = W_q[\text{MLP}(a_t) \oplus \mathcal{E}_t] \tag{3}$$

$$K_t = W_k(\mathcal{A}_t \oplus \mathcal{M}_t), \quad V_t = W_v(\mathcal{A}_t \oplus \mathcal{M}_t) \tag{4}$$

$$\text{InteractionContext}_t = \text{Softmax}\left(\frac{Q_t K_t^\top}{\sqrt{d_k}}\right) V_t \tag{5}$$

Multimodal Uncertainty Modeling: Traffic participants exhibit highly stochastic behaviors. To simulate this uncertainty, $\mathcal{D}_\theta$ does not output deterministic future coordinates but instead predicts the parameters of the positional distribution for each agent in $\mathcal{A}_{t+1}$. We employ a Gaussian Mixture Model (GMM) to fit the multimodal future:

$$P(a_{j,t+1}|s_t, a_t) = \sum_{k=1}^{K_{\text{mix}}} \pi_{j,k} \mathcal{N}(\mu_{j,k}, \Sigma_{j,k}) \tag{6}$$

where K_{mix} is the number of mixture components, $\pi_{j,k}$ is the modality weight, and $\mu_{j,k}$ and $\Sigma_{j,k}$ are the predicted mean and covariance matrices. This probabilistic formulation allows subsequent MCTS to sample worst-case scenarios during simulation, facilitating conservative and safe decision-making.

3.3 REWARD FUNCTION AND VALUE ESTIMATION

When expanding the search tree in the latent space, a signal is required to evaluate the quality of each simulated trajectory. ReVAD incorporates an immediate reward function $\mathcal{R}(s_t, a_t)$ and a long-term value network $\mathcal{V}_\phi(s_t)$.

Analytic Collision and Compliance Rewards: Benefiting from the geometric interpretability of the token space, we bypass the need for ambiguous image feature classifiers used in pixel-level models. Instead, we define analytic immediate rewards directly in the token space:

$$r_t = \omega_{\text{safe}} \mathcal{R}_{\text{coll}}(s_t) + \omega_{\text{lane}} \mathcal{R}_{\text{lane}}(s_t) + \omega_{\text{prog}} \mathcal{R}_{\text{prog}}(s_t, a_t) \tag{7}$$

Crucially, the collision penalty $\mathcal{R}_{\text{coll}}$ can be explicitly computed via the Intersection over Union (IoU) between the ego bounding box and all dynamic agent bounding boxes. If an overlap occurs, a devastating penalty is applied to instantly truncate that search branch:

$$\mathcal{R}_{\text{coll}}(s_t) = \begin{cases} -100, & \text{if } \exists a_j \in \mathcal{A}_t \text{ s.t. IoU}(\mathcal{E}_t, a_j) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

Value Network ($\mathcal{V}_\phi$): To truncate trees of infinite depth, the value network $\mathcal{V}_\phi : \mathcal{S} \to \mathbb{R}$ estimates the expected discounted return from a given state s_t:

$$\mathcal{V}_\phi(s_t) \approx \mathbb{E}_{\pi,P}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} | s_t\right] \tag{9}$$

This network consists of a Multi-Layer Perceptron (MLP) that takes the aforementioned $\text{InteractionContext}_t$ as input and outputs a scalar value.

3.4 TWO-STAGE TRAINING PARADIGM

In safety-critical domains like autonomous driving, it is unacceptable for an agent to learn models and values via trial-and-error in real-world environments (Online RL). Therefore, we propose a two-stage training paradigm completely decoupled from online interactions, utilizing only offline expert demonstration datasets $\mathcal{D}_{\text{expert}} = \{(o_t, s_t, a_t, s_{t+1}, r_t)\}$.

Stage 1: Representation and Prior Policy Learning. In the first stage, we follow the conventional E2E paradigm to train the visual encoder $\mathcal{H}$ and the imitation learning prior policy $\mathcal{P}_\psi(a|s_t)$. The policy network $\mathcal{P}_\psi$ also outputs a multimodal distribution to approximate expert behavior. The loss function $\mathcal{L}_{\text{prior}}$ is the Negative Log-Likelihood (NLL) of the expert actions:

$$\mathcal{L}_{\text{prior}}(\psi) = \mathbb{E}_{(s_t, a_t^*) \sim \mathcal{D}_{\text{expert}}}[-\log \mathcal{P}_\psi(a_t^*|s_t)] \tag{10}$$

Following Stage 1, $\mathcal{P}_\psi$ acts as a "System 1" intuitive model, serving to drastically narrow down the continuous action space during the subsequent search phase.

Stage 2: Dynamics and Value Alignment. In the second stage, we freeze the perception encoder $\mathcal{H}$ and focus on training the dynamics model $\mathcal{D}_\theta$ and the value network $\mathcal{V}_\phi$. For the dynamics model, we employ a self-supervised future prediction objective. Given the ground-truth historical state s_t and expert action a_t^*, the model must reconstruct the tokens at time $t+1$. Since dynamic agents are modeled as GMMs, the loss is the NLL of the mixture distribution; for discrete categorical attributes (e.g., traffic light states or object classes), Cross-Entropy (CE) loss is used:

$$\mathcal{L}_{\text{dyn}}(\theta) = \mathbb{E}\Bigg[-\log\left(\sum_{k=1}^{K_{\text{mix}}} \pi_k \mathcal{N}(a_{t+1}^*|\mu_k, \Sigma_k)\right) + \lambda_{\text{cls}} \text{CE}(c_{t+1}^*, \hat{c}_{t+1})\Bigg] \tag{11}$$

For the value network, we utilize Temporal Difference (TD) learning. Assuming expert trajectories are generally successful and safe, we compute the immediate reward r_t at

each step using the aforementioned analytic functions and optimize the TD(n) target:

$$z_t = \sum_{k=0}^{n-1} \gamma^k r_{t+k} + \gamma^n \mathcal{V}_{\phi_{\text{target}}}(s_{t+n}) \tag{12}$$

$$\mathcal{L}_{\text{value}}(\phi) = \mathbb{E}\left[(\mathcal{V}_\phi(s_t) - z_t)^2\right] \tag{13}$$

This purely offline two-stage alignment not only guarantees training stability but also provides the model with a robust in-distribution representational foundation.

3.5 LATENT MCTS

During deployment (Inference), ReVAD discards pure imitation execution and activates "System 2" for deliberative reasoning. Building upon the MuZero architecture, we develop Latent MCTS tailored for driving scenarios characterized by continuous action spaces and high uncertainty.

At each decision time step, MCTS constructs a search tree originating from the root node s_0 (the token state inferred in real-time from images). Each edge (s, a) in the tree stores a set of statistics: visit count $N(s, a)$, total action value $W(s, a)$, mean action value $Q(s, a)$, and prior probability $P(s, a)$. The entire search process recursively executes the following four phases for N_{sim} iterations:

1. Selection: The search traverses the tree from the root until it reaches an incompletely expanded leaf node s_L. To balance the exploitation of high-value paths and the exploration of under-visited regions during internal tree traversal, we employ a variant of the Polynomial Upper Confidence Trees (PUCT) algorithm. Specifically, given the continuous nature of the driving action space, we utilize the prior policy $\mathcal{P}_\psi(a|s)$ trained in Stage 1 to guide the search direction:

$$a^* = \arg\max_a \Bigg[Q(s,a) \quad + c_1 \cdot \mathcal{P}_\psi(a|s) \frac{\sqrt{\sum_b N(s,b)}}{1 + N(s,a)} \cdot \left(c_2 + \log \frac{\sum_b N(s,b) + c_2 + 1}{c_2} \right) \Bigg] \tag{14}$$

where c_1 and c_2 control the exploration decay rate. This mechanism ensures that, lacking sufficient simulation data, the agent leans towards "human-like" prior behaviors, but rapidly switches branches if simulations prove those behaviors dangerous.

2. Expansion & Simulation: Upon reaching a leaf node s_L, traversing a continuous action space is impossible. Thus, we sample K high-probability actions from the prior distribution $\mathcal{P}_\psi(a|s_L)$ to expand the node, initializing their statistics to $N = 0, W = 0$. Subsequently, for the selected exploratory action a, we input it into the Token Dynamics Model to simulate one step into the future:

$$s_{L+1} \sim \mathcal{D}_\theta(s_L, a) \tag{15}$$

Note: During inference, the distribution output by the GMM is either instantiated as the most probable mean to prevent tree branching explosion, or sampled via Monte Carlo with a temperature parameter to evaluate extreme corner cases.

3. Evaluation: The newly generated latent state s_{L+1} is evaluated using the reward function and the value network. ReVAD's core advantage shines here: if a collision is detected in the token geometry of s_{L+1} (i.e., $\mathcal{R}_{\text{coll}} < 0$), the node's value v is immediately overwritten with a massive negative penalty, permanently terminating further search down this hazardous branch. Otherwise, $v = \mathcal{V}_\phi(s_{L+1})$.

4. Backpropagation: The cumulative discounted return for the leaf node, $G = r_L + \gamma \cdot v$, is computed and backpropagated up the search path to the root, updating the statistics of all traversed edges (s, a):

$$N(s,a) \leftarrow N(s,a) + 1 \tag{16}$$

$$W(s,a) \leftarrow W(s,a) + G \tag{17}$$

$$Q(s,a) = \frac{W(s,a)}{N(s,a)} \tag{18}$$

5. Action Selection: After completing N_{sim} iterations, ReVAD aggregates the visit counts of all candidate actions at the root node s_0. To maximize robustness against OOD scenarios, the system selects the action a_{final} most frequently visited by MCTS (i.e., the action subjected to the most safety verifications) for execution by the vehicle chassis:

$$a_{\text{final}} = \arg\max_a N(s_0, a) \tag{19}$$

Through this comprehensive System 2 paradigm, ReVAD achieves real-time counterfactual evaluation within an exceptionally lightweight token space, fundamentally overcoming the "blind overconfidence" flaw inherent to pure imitation learning in long-tail scenarios.

4 EXPERIMENTS

To rigorously demonstrate the superiority of the deliberative "System 2" reasoning in ReVAD over reactive "System 1" baselines, we conduct comprehensive evaluations across high-fidelity simulators and large-scale real-world datasets. Our experiments are designed to address the following critical aspects of autonomous driving: 1) The upper bound of closed-loop driving performance on standard benchmarks. 2) The notoriously known open-loop vs. closed-loop evaluation gap. 3) Robustness in OOD events and scalability in ultra-dense traffic. 4) Spatiotemporal kinematic consistency and ride comfort. 5) Detailed computational overhead and the internal decision-making mechanism of Latent MCTS.

4.1 EXPERIMENTAL SETTINGS

Datasets and Benchmarks. Following the rigorous evaluation protocol of VADv2, we assess ReVAD on three authoritative benchmarks:

- **CARLA Benchmarks Jia et al. [2024]:** We evaluate closed-loop planning on the Town05 Long and Bench2Drive benchmarks. The primary metrics include Route Completion (RC), Infraction Score (IS), and Driving Score (DS). Simulation runs at 10 Hz.
- **NAVSIM Dataset Dauner et al. [2024]:** A large-scale real-world dataset evaluated using the official Predictive Driving Metric Score (PDMS) and No Collision (NC) rate.
- **3DGS-based Benchmark Kerbl et al. [2023]:** A photorealistic closed-loop environment containing 337 reconstructed dense traffic scenarios. Metrics include Collision Ratio (CR) and Deviation Ratio (DR).

Implementation Details. For a strictly fair algorithmic comparison, ReVAD adopts the exact same visual backbone (ResNet-50) and Transformer-based representation encoder $\mathcal{H}$ as VADv2. Furthermore, the pre-trained VADv2 model directly serves as our Imitation Prior Policy $\mathcal{P}_\psi$. For the Latent MCTS, the default hyperparameters are defined as follows: simulation budget $N_{\text{sim}} = 50$, expansion candidates $K = 5$, search depth $H_s = 3$ steps (evaluating 1.5 seconds into the future at 0.5s intervals), and discount factor $\gamma = 0.95$. The probabilistic dynamics model $\mathcal{D}_\theta$ is trained offline on the 3 million clips collected from CARLA without any online reinforcement learning. All models are deployed and evaluated on a single NVIDIA RTX 4090 GPU.

4.2 MAIN RESULTS ON STANDARD BENCHMARKS

Closed-loop Performance on CARLA. As reported in Table 1, ReVAD achieves SOTA closed-loop performance on the Town05 Long benchmark. While VADv2 demonstrates strong performance (DS: 85.1) through probabilistic regression, ReVAD elevates the Driving Score to an unprecedented 89.2. Notably, the Infraction Score (IS) drastically increases from 0.87 to 0.93. This specific improvement is theoretically significant: pure generative models often suffer from "hallucinated compliance"—they may visually encode a red light but fail to ground it geometrically, leading to marginal infractions like line-crossing or stop-sign running due to statistical noise. By explicitly rolling out future states in the latent space and applying hard geometric penalty checks, ReVAD acts as an algorithmic safeguard, perfectly aligning the agent's behavior with strict traffic rules and preventing the execution of hazardous "human-like" but contextually flawed prior actions. On the more demanding Bench2Drive benchmark, ReVAD identically establishes a new SOTA with a DS of **79.4**.

Table 1: Closed-loop evaluation on the CARLA Town05 Long benchmark. ReVAD drastically reduces infractions, achieving the highest Driving Score (DS).

Method	Modality	DS ↑	RC ↑	IS ↑
Transfuser Prakash et al. [2021]	C+L	31.0	47.5	0.77
DriveMLM Cui et al. [2025]	C+L	76.1	98.1	0.78
MILE Hu et al. [2022]	C	61.1	97.4	0.63
VAD Jiang et al. [2023]	C	30.3	75.2	-
DriveCoT Wang et al. [2024a]	C	73.6	96.8	0.76
LeapVAD Ma et al. [2026]	C	73.7	95.7	0.78
VADv2 (System 1 Prior)	C	85.1	98.4	0.87
ReVAD (Ours, System 2)	C	**89.2**	**98.6**	**0.93**

Generalization on Real-world Datasets. To verify that ReVAD's deliberative reasoning transfers seamlessly to real-world complexities, we evaluate it on the NAVSIM and 3DGS benchmarks (Table 2). In NAVSIM, ReVAD achieves a leading PDMS of 91.5. However, static log-replay datasets like NAVSIM cannot measure how an agent's action perturbs the environment. Therefore, the 3DGS-based benchmark, which involves reactive agents, provides a truer test of closed-loop robustness. Here, ReVAD dramatically reduces the Collision Ratio (CR) to 0.185—a remarkable 31.4% relative reduction compared to the base VADv2 (0.270). Pure behavioral cloning models suffer from "causal compounding errors" in interactive settings because they cannot anticipate how other vehicles will react to their maneuvers. Conversely, ReVAD's internal dynamics model explicitly simulates this multi-agent interaction, empowering the Latent MCTS to proactively navigate around uncooperative obstacles before a critical conflict manifests.

Table 2: Performance on Real-world Benchmarks. ReVAD demonstrates exceptional safety in dense traffic scenarios, drastically reducing the 3DGS Collision Ratio.

Method	NAVSIM (navtest)		3DGS Benchmark	
	PDMS ↑	NC ↑	CR ↓	DR ↓
UniAD Hu et al. [2023b]	83.4	97.8	-	-
DiffusionDrive Liao et al. [2025]	88.1	98.2	-	-
GenAD Zheng et al. [2024]	-	-	0.341	0.291
VAD Jiang et al. [2023]	-	-	0.335	0.314
VADv2 (Prior)	89.3	98.3	0.270	**0.243**
ReVAD (Ours)	**91.5**	**98.8**	**0.185**	0.248

4.3 BRIDGING THE OPEN-LOOP VS. CLOSED-LOOP GAP

A fundamental dilemma in E2E driving is the severe disconnect between open-loop metrics (e.g., L2 displacement error against expert trajectories) and closed-loop survival capabilities. Imitation learning models are explicitly optimized for

open-loop likelihood, often leading to compounding errors during closed-loop autoregressive execution.

We compare VADv2 and ReVAD under both settings (Table 3). Interestingly, VADv2 achieves a marginally lower open-loop L2 error at 3s (0.225m vs. 0.231m for ReVAD). This marginal discrepancy occurs because ReVAD's MCTS intentionally deviates from the "expert average" when its forward simulation detects a potential latent collision. While this conservative deviation slightly penalizes open-loop alignment, it yields massive dividends in closed-loop deployment, dropping the collision rate from 0.007% to 0.002%. This confirms that optimizing for trajectory likelihood (System 1) does not inherently optimize for survival (System 2); explicit causal reasoning is indispensable for safe autonomy.

Table 3: Open-Loop vs. Closed-Loop Gap. ReVAD sacrifices negligible open-loop L2 accuracy to achieve a massive reduction in closed-loop collisions.

Method	Open-Loop		Closed-Loop
	L2 @ 1s ↓	L2 @ 3s ↓	Collision ↓
VADv2 (Prior)	**0.082**	**0.225**	0.007%
ReVAD (Ours)	0.084	0.231	**0.002%**

4.4 ROBUSTNESS IN HIGH-UNCERTAINTY AND DENSITY SCENARIOS

Adversarial OOD Stress Test. The fatal flaw of probabilistic imitation is its reliance on the training manifold. Without a transition model, agents cannot foresee the consequences of their actions in OOD scenarios. We engineered an Adversarial Interactive Stress Test in CARLA, featuring rare, high-uncertainty events (e.g., sudden pedestrian emergence from occlusion, and extreme aggressive vehicle cut-ins).

As shown in Table 4, under standard traffic, both methods perform exceptionally well. However, under aggressive cut-ins, VADv2's Collision Avoidance Success Rate (CASR) collapses to 42.5%. It blindly executes its maximum-likelihood prediction, failing to realize the environmental dynamics have drifted. Conversely, ReVAD maintains a robust CASR of 81.2%. By leveraging $\mathcal{D}_\theta$ to simulate the bounding-box overlaps resulting from the cut-in, MCTS assigns massive penalties to the prior's trajectory, aggressively shifting the decision tree towards emergency braking or evasive steering.

Scalability with Traffic Density. Multi-agent interaction complexity scales exponentially with traffic density. We categorize driving scenes in Town05 based on the number of dynamic agents within a 30-meter radius: Low (<5), Medium (5-15), and High (>15). Table 5 illustrates the scalability of different planners. VADv2's DS drops significantly by 14.3 points from Low to High density. In contrast, ReVAD exhibits graceful degradation, with DS dropping by only 8.1 points. The cross-attention mechanism within our Dynamics Model effectively captures the joint probability of multi-agent interactions, whereas baseline methods treat marginal distributions of agents independently, leading to paralysis in dense traffic.

Table 4: Adversarial OOD Stress Test: Collision Avoidance Success Rate (CASR ↑). ReVAD exhibits supreme robustness against extreme corner cases.

Method	Standard Traffic	Sudden Occlusion	Aggressive Cut-in
Transfuser	78.4%	22.1%	15.3%
VADv2 (System 1)	96.2%	68.4%	42.5%
ReVAD (System 2)	**98.5%**	**89.6%**	**81.2%**

Table 5: Scalability with Traffic Density (Driving Score ↑).

Method	Low (<5)	Medium (5-15)	High (>15)
DriveMLM	82.4	75.3	61.2
VADv2 (Prior)	92.1	85.5	77.8
ReVAD (Ours)	**93.5**	**89.2**	**85.4**

4.5 KINEMATIC CONSISTENCY AND RIDE COMFORT

A common issue in generative and probabilistic driving models is "trajectory flickering"—where the predicted multimodal outputs rapidly oscillate between frames, causing severe passenger discomfort. By aggregating future values over multiple MCTS simulations, ReVAD intrinsically acts as a temporal smoothing filter. Table 6 evaluates ride comfort using Maximum Jerk, Maximum Lateral Acceleration, and Trajectory Flickering Rate (TFR, defined as the percentage of frames where the planned lateral intent flips). ReVAD reduces TFR by over 50%, resulting in a profoundly smoother and more stable physical execution.

Table 6: Kinematic Consistency and Ride Comfort Metrics.

Method	Max Jerk ↓ (m/s^3)	Max Lat. Accel ↓ (m/s^2)	Flickering Rate ↓ (TFR %)
Transfuser	4.12	3.25	18.4%
VADv2 (Prior)	2.65	2.10	9.2%
ReVAD (Ours)	**1.82**	**1.45**	**4.1%**

4.6 COMPREHENSIVE ABLATION STUDIES

MCTS Components and Value Network. Table 7 isolates the contribution of each algorithmic component. Removing the explicit collision penalty ($\mathcal{R}_{\text{coll}}$) causes the IS to drop sharply from 0.93 to 0.88, proving that analytic geometry checking in the token space is the primary driver of safety. If we remove the TD-trained Value Network $\mathcal{V}_\phi$

and rely solely on immediate rewards, the agent becomes myopic, reducing DS to 86.4. Furthermore, replacing the informed Imitation Prior ($\mathcal{P}_\psi$) with a uniform random expansion plummets DS to 61.4, highlighting the indispensable synergy: System 1 provides computationally tractable "human-like" proposals, while System 2 ensures rigorous safety bounds.

Table 7: Ablation on Latent MCTS Components (Town05 Long).

Configuration	DS ↑	RC ↑	IS ↑
Full ReVAD (Latent MCTS)	**89.2**	**98.6**	**0.93**
w/o Explicit Collision Penalty ($\mathcal{R}_{\text{coll}}$)	84.7	98.5	0.88
w/o Value Network (Myopic Search)	86.4	97.2	0.89
w/o Imitation Prior (Uniform Search)	61.4	82.3	0.74

Search Depth (H_s). Table 8 demonstrates the trade-off between search depth H_s and safety. A depth of $H_s = 1$ (looking ahead 0.5s) is insufficient to prevent high-speed collisions. Increasing depth to $H_s = 3$ (1.5s) optimizes safety, capturing enough causal dynamics to successfully override dangerous priors.

Table 8: Ablation on MCTS Search Depth (H_s).

Search Depth (H_s)	Lookahead Time	Infraction Score ↑
0 (VADv2 Baseline)	0.0 s	0.87
1	0.5 s	0.89
2	1.0 s	0.91
3 (Default)	**1.5 s**	**0.93**
4	2.0 s	0.93

4.7 COMPUTATIONAL OVERHEAD BREAKDOWN

A common critique of tree-search algorithms is their computational overhead, which often precludes real-time deployment. Because ReVAD operates exclusively in a highly sparse Token Space (requiring only lightweight Transformer matrix multiplications) rather than rendering dense pixels, it remains exceptionally efficient. Table 9 profiles the latency of each module per frame. Executing 50 MCTS simulations introduces merely 17 ms of overhead (Dynamics + Value + Backprop) compared to the base VADv2 architecture. The total E2E latency is 142 ms, safely satisfying the $\sim$10 Hz real-time control constraints required by autonomous chassis systems.

4.8 QUANTITATIVE SCENARIO-BASED DECISION ANALYSIS

To intuitively and quantitatively demystify the cognitive override process of ReVAD, we present a Scenario-based Decision Analysis in Table 10. This table tracks the internal metrics of three highly critical driving scenarios, comparing the System 1 imitation probability ($\mathcal{P}_\psi$) against the System 2 estimated action-value ($Q(s, a)$) after MCTS rollout.

Table 9: E2E Inference Latency Breakdown ($N_{\text{sim}} = 50, H_s = 3$).

Module (System 1)	Time	Module (System 2 MCTS)	Time
Visual Perception (BEV)	35 ms	Dynamics Sim ($\mathcal{D}_\theta \times 50$)	9 ms
Tokenization ($\mathcal{H}$)	42 ms	Value Evaluation ($\mathcal{V}_\phi \times 50$)	6 ms
Prior Policy ($\mathcal{P}_\psi$)	48 ms	Tree Expansion & Backprop	2 ms
Subtotal (Prior)	**125 ms**	**Subtotal (MCTS Overhead)**	**17 ms**
Total ReVAD Inference Latency per frame			**142 ms**

Table 10: Quantitative Scenario-based Decision Analysis. MCTS explicitly penalizes high-probability but hazardous imitation actions, forcing safe execution.

Driving Scenario	Candidate Action	Sys 1 Prob ($\mathcal{P}_\psi$)	Sys 2 Value (Q)	Final Execution
1. Unprotected Left Turn	Action A: *Aggressive Crossing*	**0.82** (High)	-95.4 (Collision)	**Rejected**
	Action B: *Yield to Oncoming*	0.12 (Low)	**+12.5 (Safe)**	Selected
2. High-speed Cut-in	Action A: *Maintain Speed*	**0.75** (High)	-88.0 (Collision)	**Rejected**
	Action B: *Emergency Brake*	0.08 (Low)	**+8.4 (Safe)**	Selected
3. Pedestrian Emergence	Action A: *Swerve Left*	**0.55** (High)	-45.2 (Lane Viol.)	**Rejected**
	Action B: *Hard Stop*	0.25 (Med)	**+9.2 (Safe)**	Selected

Case Study: Unprotected Left Turn. In Scenario 1, the ego-vehicle turns left while a high-speed oncoming vehicle approaches. The System 1 imitation prior, biased by dataset statistics (where oncoming vehicles often yield), assigns 82% probability to Action A (Aggressive Crossing). Executed blindly, this results in a severe T-bone collision. ReVAD instead expands the node with System 2 reasoning. The Token Dynamics Model predicts a multimodal future for the oncoming vehicle, including a worst-case mode where it maintains speed, and the rollout detects an IoU overlap between ego and oncoming agent tokens. MCTS then back-propagates a catastrophic penalty ($Q = -95.4$) to the root, causing PUCT to abandon this branch and explore Action B (Yield). With collision-free rollouts ($Q = +12.5$), ReVAD executes the low-probability yet safe yielding action, exemplifying "thinking before acting."

5 CONCLUSION

In this paper, we present ReVAD to bridge the gap between reactive imitation and deliberative planning in E2E autonomous driving. ReVAD leverages a pre-trained imitation policy as a "System 1" prior and employs a probabilistic Token Dynamics Model with Latent MCTS for "System 2" counterfactual reasoning. Extensive evaluations show that explicitly rolling out future dynamics in the token space drastically reduces closed-loop collisions and mitigates catastrophic failures in OOD scenarios. Furthermore, our highly optimized MCTS introduces minimal latency, proving the practical viability of real-time reasoning for safe autonomy.

Acknowledgements

This work was supported by the National Key R&D Program of China (No. 2024YFB4303400), the Key Science and Technology Program of Shaanxi Province, China (Grant No. 2025CY-YBXM-197), and the Natural Science Basic Research Plan in Shaanxi Province of China (Program No. 2025JC-YBQN-811).

References

Felipe Codevilla, Eder Santana, Antonio M López, and Adrien Gaidon. Exploring the limitations of behavior cloning for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9329–9338, 2019.

Erfei Cui, Wenhai Wang, Zhiqi Li, Jiangwei Xie, Haoming Zou, Hanming Deng, Gen Luo, Lewei Lu, Xizhou Zhu, and Jifeng Dai. Drivemlm: aligning multi-modal large language models with behavioral planning states for autonomous driving. *Visual Intelligence*, 3(22), 2025.

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

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations (ICLR)*, 2020.

Shadi Hamdan, Chonghao Sima, Zetong Yang, Hongyang Li, and Fatma Güney. ETA: Efficiency through thinking ahead, a dual approach to self-driving with large models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 26529–26538, October 2025. doi: 10.1109/ICCV51701.2025.02462.

Anthony Hu, Gianluca Corrado, Nicolas Griffiths, Zachary Murez, Corina Gurau, Hudson Yeo, Alex Kendall, Roberto Cipolla, and Jamie Shotton. Model-based imitation learning for urban driving. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. GAIA-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023a.

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)*, June 2023b.

Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Bo Jiang, Shaoyu Chen, Hao Gao, Bencheng Liao, Qian Zhang, Wenyu Liu, and Xinggang Wang. Vadv2: End-to-end autonomous driving via probabilistic planning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.

Yingyan Li, Lue Fan, Jiawei He, Yuqi Wang, Yuntao Chen, Zhaoxiang Zhang, and Tieniu Tan. Enhancing end-to-end autonomous driving with latent world model. In *International Conference on Learning Representations (ICLR)*, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Yukai Ma, Tiantian Wei, Naiting Zhong, Jianbiao Mei, Tao Hu, Licheng Wen, Xuemeng Yang, Botian Shi, and Yong Liu. Leapvad: A leap in autonomous driving via cognitive perception and dual-process thinking. *IEEE Transactions on Neural Networks and Learning Systems*, 37(4):1963–1977, April 2026. doi: 10.1109/TNNLS.2025.3626711.

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

S. Momchil et al. TreeIRL: Safe urban driving with tree search and inverse reinforcement learning. *arXiv preprint arXiv:2509.13579*, 2025.

Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7077–7087, 2021.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

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

Tianqi Wang, Enze Xie, Ruihang Chu, Zhenguo Li, and Ping Luo. Drivecot: Integrating chain-of-thought reasoning with end-to-end driving. *arXiv preprint arXiv:2403.16996*, 2024a.

Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Wayve. LINGO-1: Exploring natural language for autonomous driving. Wayve article, 2023.

Wenzhao Zheng, Ruiqi Song, Xianda Guo, Chenming Zhang, and Long Chen. Genad: Generative end-to-end autonomous driving. In *Computer Vision – ECCV 2024*, volume 15123 of *Lecture Notes in Computer Science*, pages 87–104, 2024.

Yupeng Zheng, Pengxuan Yang, Zebin Xing, Qichao Zhang, Yuhang Zheng, Yinfeng Gao, Pengfei Li, Teng Zhang, Zhongpu Xia, Peng Jia, XianPeng Lang, and Dongbin Zhao. World4drive: End-to-end autonomous driving via intention-aware physical latent world model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.