ExploreVLA: Dense World Modeling and Exploration for End-to-End Autonomous Driving

Zihao Sheng[1,2]*, Xin Ye[1], Jingru Luo[1], Sikai Chen[2], and Liu Ren[1]

[1] Bosch Research North America & Bosch Center for Artificial Intelligence (BCAI)
[2] University of Wisconsin–Madison
{zihao.sheng,sikai.chen}@wisc.edu
{xin.ye3,jingru.luo,liu.ren}@us.bosch.com

Abstract. End-to-end autonomous driving models based on Vision-Language-Action (VLA) architectures have shown promising results by learning driving policies through behavior cloning on expert demonstrations. However, imitation learning inherently limits the model to replicating observed behaviors without exploring diverse driving strategies, leaving it brittle in novel or out-of-distribution scenarios. Reinforcement learning (RL) offers a natural remedy by enabling policy exploration beyond the expert distribution. Yet VLA models, typically trained on offline datasets, lack directly observable state transitions, necessitating a learned world model to anticipate action consequences. In this work, we propose a unified understanding-and-generation framework that leverages world modeling to simultaneously enable meaningful exploration and provide dense supervision. Specifically, we augment trajectory prediction with future RGB and depth image generation as dense world modeling objectives, requiring the model to learn fine-grained visual and geometric representations that substantially enrich the planning backbone. Beyond serving as a supervisory signal, the world model further acts as a source of intrinsic reward for policy exploration: its image prediction uncertainty naturally measures a trajectory's novelty relative to the training distribution, where high uncertainty indicates out-of-distribution scenarios that, if safe, represent valuable learning opportunities. We incorporate this exploration signal into a safety-gated reward and optimize the policy via Group Relative Policy Optimization (GRPO). Experiments on the NAVSIM and nuScenes benchmarks demonstrate the effectiveness of our approach, achieving a state-of-the-art PDMS score of 93.7 and an EPDMS of 88.8 on NAVSIM. The code is available at https://zihaosheng.github.io/ExploreVLA/.

Keywords: Autonomous driving · Vision-language-action model · World action model

1 Introduction

End-to-end autonomous driving has advanced rapidly with the emergence of Vision-Language-Action (VLA) architectures [20, 21, 38, 55], which unify percep-

*Work was done during internship at Bosch, supervised by Xin Ye

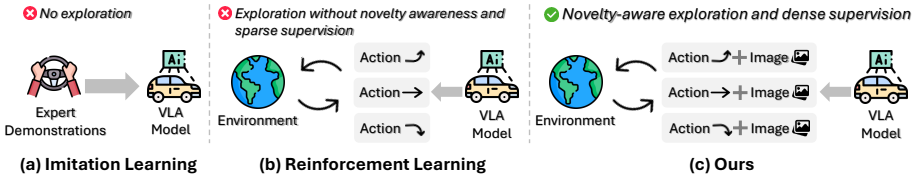

Fig. 1: Comparison of training paradigms for VLA-based autonomous driving. (a) Imitation learning directly clones expert demonstrations without exploration. (b) Previous reinforcement learning enables policy exploration, but cannot distinguish expert imitation from genuine out-of-distribution discovery and relies on sparse supervision. (c) Our approach augments RL with dense world modeling supervision via future image generation, while leveraging image prediction uncertainty as a novelty measure to identify and prioritize valuable exploratory strategies.

tion, reasoning, and planning within a single model. By leveraging the representational power of large vision-language models, these models have demonstrated promising capabilities in translating raw sensor observations into driving actions. Yet, the dominant training paradigm for such models (behavior cloning on expert demonstrations or supervised fine-tuning) introduces a fundamental bottleneck: the learned policy can only replicate the behaviors it has observed, without the ability to discover alternative strategies that may be equally or more effective. This limitation manifests as distributional brittleness: when confronted with scenarios that deviate from the expert distribution, the policy lacks the exploratory experience needed to generalize [6, 33].

Reinforcement learning (RL) offers a principled mechanism to overcome this limitation by allowing the agent to explore beyond the boundaries of expert data and optimize its policy through trial-and-error interaction [14]. Recent advances in RL post-training, exemplified by Group Relative Policy Optimization (GRPO) [34], have demonstrated that sampling diverse candidate outputs and performing relative ranking can effectively improve policy quality atop strong pretrained models. However, applying RL to autonomous driving poses challenges distinct from other domains. In language tasks, state transitions are fully determined by the model's output, and outcomes are immediately observable. In robotics, high-fidelity simulators enable safe trial-and-error. Autonomous driving enjoys neither advantage: the consequences of a planned trajectory depend on complex scene dynamics that no existing simulator faithfully captures. These challenges motivate the need for a world model that can internalize environment dynamics and anticipate action consequences from data alone. Yet a further challenge remains: standard task-level rewards such as Predictive Driver Model Score (PDMS) [7] evaluate trajectory quality but do not distinguish between policies that merely replicate expert behavior and those that have genuinely discovered novel strategies. An exploration signal orthogonal to task performance is therefore essential to unlock the full potential of RL post-training.

A second limitation of existing VLA driving models is the sparsity of their supervisory signals. Most approaches rely on textual descriptions and trajectory waypoints as training targets [55, 56], which, while informative for high-level decision-making, fail to capture the rich spatial geometry and fine-grained appearance of driving scenes. This supervisory deficit constrains the model's ability to build comprehensive representations, particularly for aspects of the environment (such as road topology, object extent, and depth ordering) that are critical for safe planning but are not explicitly encoded in sparse action labels [27].

In this work, we propose a unified understanding-and-generation framework that addresses both limitations through a single mechanism: **dense world modeling and exploration** (Fig. 1). Specifically, we augment trajectory prediction with future RGB and depth image generation as auxiliary objectives. On the supervision side, these generation tasks require the model to predict fine-grained visual appearance and metric geometry of future scenes, providing dense gradient signals that substantially enrich the planning backbone's visual and geometric representations. On the exploration side, we leverage a key insight: *the world model's image prediction uncertainty naturally measures a trajectory's novelty relative to the training distribution*. Since the world model is trained exclusively on expert demonstrations, it produces low-uncertainty predictions for trajectories within the expert distribution but exhibits high uncertainty on out-of-distribution (OOD) trajectories. Crucially, this uncertainty reflects distance to the *entire* training distribution, rather than deviation from a single ground-truth trajectory, which provides a more reliable novelty signal than trajectory distance alone. We incorporate this signal into a safety-gated exploration reward: trajectories that achieve high PDMS scores while exhibiting high prediction uncertainty are identified as valuable discoveries of new successful strategies and receive an exploration bonus. This composite reward is optimized via GRPO, enabling the policy to expand its behavioral repertoire while maintaining driving safety.

Our contributions can be summarized as follows:

- We introduce a novel exploration mechanism for RL post-training that uses the world model's image prediction uncertainty as an intrinsic novelty measure, coupled with a safety-gated reward to encourage beneficial out-of-distribution exploration.
- We propose a unified VLA framework that jointly predicts future trajectories, RGB images, and depth images, leveraging dense world modeling to provide rich visual and geometric supervision for the planning backbone.
- We demonstrate state-of-the-art performance on the NAVSIM benchmark with a PDMS of 93.7 and an EPDMS of 88.8, and validate the generalizability of our approach on the nuScenes dataset.

2 Related Work

2.1 World Models for Autonomous Driving

In the context of autonomous driving, world models offer the ability to predict future states of the driving scene, enabling planning, data augmentation, and

closed-loop evaluation without costly real-world interactions [8, 10, 13, 42]. A significant line of work focuses on video prediction-based world models [12, 15, 43]. For example, GAIA-1 [16] leverages a generative model conditioned on video, text, and action inputs to produce realistic driving videos. DriveDreamer [36] generates future driving frames conditioned on HDMaps and 3D bounding boxes. GenAD [44] proposes an action-conditioned video generation framework that supports long-horizon future prediction for planning. Another prominent direction explores world models within structured geometric spaces. UniWorld [32] proposes a unified framework for 4D occupancy forecasting, and OccWorld [53] extends this idea by predicting 3D occupancy and ego-motion jointly, providing a compact world representation for downstream planning. UniScene [24] further scales this paradigm by jointly generating consistent future semantic occupancy, LiDAR and multi-view images. Beyond generation and prediction, several works integrate world models into the planning pipeline. For instance, WoTE [28] incorporates a BEV world model to predict future states for online trajectory evaluation and selection. OmniNWM [25] employs a learned navigation world model to generate imagined futures and derive planning rewards from predicted scene dynamics. World4Drive [54] further introduces a latent world model that predicts future latent states under multiple driving intentions and selects trajectories via a learned selector. In our work, we leverage the world model not only for future state prediction but also as a source of intrinsic reward signals that encourage the policy to explore diverse and informative driving behaviors, thereby improving generalization beyond the coverage of the training data.

2.2 VLA Models for Autonomous Driving

The integration of vision, language, and action within a unified framework has emerged as a promising paradigm for autonomous driving [23, 29]. Early efforts, such as DriveGPT-4 [41], use frozen VLMs to narrate driving scenes but do not directly output control signals. Subsequent modular VLA approaches began embedding language into the planning loop. For example, OpenDriveVLA [55] fuses multimodal sensor inputs with textual route instructions to generate interpretable waypoints, while RAG-Driver [49] introduces retrieval-augmented planning for long-tail scenarios. The field then advances toward unified end-to-end architectures. EMMA [20] jointly performs detection and planning within a single VLM, and DiffVLA [21] combines diffusion-based trajectory sampling with language-conditioned embeddings. Most recently, reasoning-augmented VLA models have pushed the frontier further. ORION [11] incorporates a transformer memory module for long-horizon reasoning, and AutoVLA [56] fuses CoT reasoning and trajectory planning in a single autoregressive transformer. Alpamayo-R1 [38] introduces causally grounded Chain-of-Causation reasoning tightly integrated with trajectory prediction, enhancing reasoning-action consistency and long-tail safety performance. Despite this rapid progress, most existing VLA models rely on textual descriptions and action trajectories as the primary supervisory signal, which are inherently sparse. This supervisory sparsity limits the model's ability to learn comprehensive scene representations. In contrast,

our work leverages RGB and depth images as auxiliary dense supervisory signals to encourage the model to capture richer visual and geometric cues, thereby yielding more accurate and robust trajectory planning.

2.3 Unified Understanding and Generation Models

In foundation model research, the long-standing separation between understanding and generation has motivated a growing effort to unify both within a single architecture for greater scalability and cross-task synergy [3, 37, 39]. In autonomous driving, this paradigm has gained significant traction as researchers seek to bridge the gap between world modeling and end-to-end planning. For example, FutureSightDrive [50] proposes a spatio-temporal visual Chain-of-Thought framework where a VLA model first generates future frames, including lane lines, 3D bounding boxes, and complete future images, as visual intermediate reasoning steps, then predicts trajectories conditioned on these imagined futures. Policy World Model (PWM) [52] pre-trains on large-scale action-free video generation to learn world dynamics, then fine-tunes with a collaborative formulation where trajectory planning is explicitly conditioned on forecasted future states. DriveVLA-W0 [27] introduces world modeling objectives to unlock data scaling laws for end-to-end driving. UniDrive-WM [40] unifies scene understanding, trajectory planning, and trajectory-conditioned future image generation within a single VLM. Epona [51] combines autoregressive causal modeling with diffusion-based generation through decoupled spatiotemporal factorization, enabling both high-fidelity video synthesis and trajectory planning. Together, these works demonstrate that jointly modeling future scene generation and action prediction yields more informed and anticipatory planning. Our work follows this unified paradigm by leveraging RGB and depth image generation as dense supervisory signals alongside trajectory prediction, while further employing the world model to provide intrinsic reward signals that encourage exploratory driving behaviors and improve generalization beyond the training distribution.

3 Methodology

3.1 Overview

We present a unified understanding-and-generation framework for end-to-end autonomous driving that addresses two key limitations of existing VLA models: (1) the lack of exploration beyond expert demonstrations and (2) the reliance on sparse supervisory signals. Our framework consists of three components. First, we build upon a unified VLM backbone that jointly supports autoregressive text modeling and discrete image generation within a single architecture (Sec. 3.3). Second, we introduce future RGB and depth image generation as dense world modeling objectives that provide token-level supervision alongside trajectory prediction, encouraging the model to learn richer scene representations (Sec. 3.4). Third, we leverage the world model's prediction uncertainty as an intrinsic reward signal to guide policy exploration via Group Relative Policy Optimization

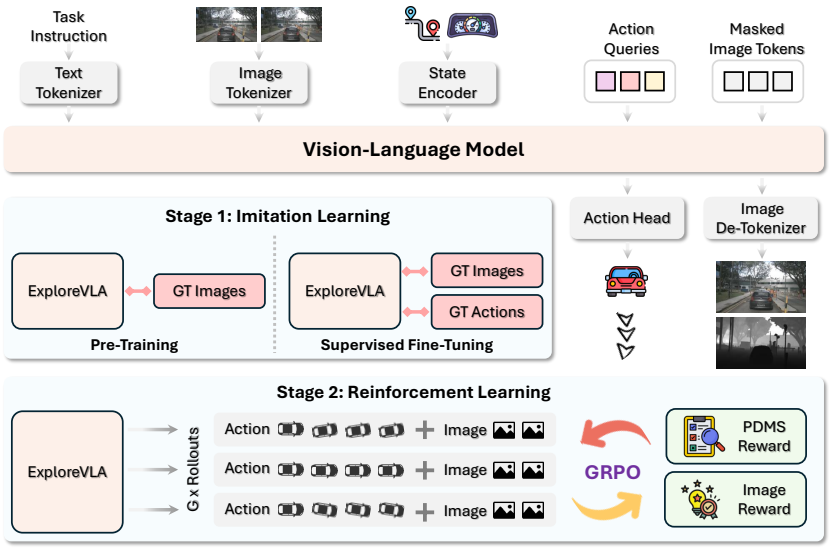

Fig. 2: Model architecture and training paradigm of ExploreVLA. The model takes task instructions, multi-frame images, and ego status as input, and jointly predicts future trajectories and future images. Training proceeds in two stages: (1) imitation learning, consisting of pre-training on image generation and supervised fine-tuning on both actions and images, and (2) reinforcement learning, where GRPO optimizes the policy using a composite reward combining PDMS and image-based exploration bonus.

(GRPO), enabling the model to discover diverse driving strategies beyond mere imitation (Sec. 3.5). An overview of our framework is illustrated in Fig. 2.

3.2 Problem Formulation

We formulate autonomous driving as a unified understanding-and-generation task, where a single model jointly predicts future trajectories and generates dense visual representations conditioned on the current driving context.

Input. At each timestep t, the model receives three inputs: (1) the current and T past front-view camera images $\{\mathbf{I}_{t-T}, \cdots, \mathbf{I}_t\}$ with $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, (2) a natural language command $\mathbf{c}$, and (3) the ego-vehicle status $\mathbf{s}_t$.

Output. The model produces three types of outputs: (1) predicted future waypoints $\boldsymbol{\tau} = \{\hat{p}_i\}_{i=1}^{N_\tau}$, where N_τ denotes the planning horizon, (2) predicted future RGB frames $\{\hat{\mathbf{I}}_{t+1}, \cdots, \hat{\mathbf{I}}_{t+F}\}$ with $\hat{\mathbf{I}} \in \mathbb{R}^{H' \times W' \times 3}$ capturing the anticipated visual appearance of the driving scene, and (3) predicted future depth maps $\{\hat{\mathbf{D}}_{t+1}, \cdots, \hat{\mathbf{D}}_{t+F}\}$ with $\hat{\mathbf{D}} \in \mathbb{R}^{H' \times W'}$ encoding the geometric structure of the future scene. Here F denotes the number of future frames to generate.

3.3 Unified Understanding and Generation Architecture

Tokenization. We maintain a unified vocabulary of discrete tokens spanning text, images, and special task indicators. For **text tokenization**, we adopt the same tokenizer from the pre-trained LLM backbone, such that the language command $\mathbf{c}$ is tokenized into L text tokens $\mathbf{v} = \{v_1, v_2, \cdots, v_L\}$. For **image tokenization**, we employ a pre-trained MAGVIT-v2 [47] quantizer with a lookup-free codebook of size $K = 8{,}192$. Each image is encoded into discrete tokens by partitioning it into non-overlapping patches of size 16×16. Each of the $(T+1)$ input frames is independently tokenized into M image tokens, yielding the input image token sequence $\mathbf{u} = \{u_1, u_2, \cdots, u_{(T+1)\times M}\}$. For **ego status encoding**, the $\mathbf{s}_t$ is projected into the transformer's embedding space via a learnable MLP, producing ego status embeddings $\mathbf{e}_s$ that are concatenated with the other input tokens.

Causal and Full Attention Mechanism. We adopt the omni-attention mechanism from Show-o [39], which adaptively combines causal and full attention depending on the token type. Text tokens $\mathbf{v}$ and ego status embeddings $\mathbf{e}_s$ are processed via *causal attention*, where each token attends only to its preceding tokens. Image tokens $\mathbf{u}$ are processed via *full attention*, which allows comprehensive interaction among all spatial positions and ensures that the generation process is fully conditioned on the driving context.

Trajectory Prediction Head. In addition to the generation outputs, we extract the hidden states from the transformer at designated positions and feed them through a lightweight MLP head to predict future waypoints:

$$\boldsymbol{\tau} = \mathrm{MLP}(\mathbf{h}), \tag{1}$$

where $\mathbf{h}$ denotes the hidden representation. This design decouples trajectory prediction from the discrete token vocabulary, allowing the model to produce continuous-valued waypoints while sharing the same contextualized representations learned through the joint understanding-and-generation objective.

3.4 Dense Supervisory Signals via World Modeling

A central limitation of existing VLA models for autonomous driving is their reliance on sparse supervisory signals. Textual descriptions provide only high-level semantic intent (*e.g.*, "turn left"), while trajectory waypoints encode a thin, low-dimensional slice of the rich information embedded in driving scenes. As a result, the vast majority of scene structure, such as patial layout and depth ordering, remains unsupervised, which limits the model's ability to learn comprehensive representations of the driving environment.

We address this by formulating future scene generation as an auxiliary world modeling objective that provides dense supervision alongside trajectory prediction. Our model is trained to generate F future frames of both RGB images and

depth maps, effectively requiring it to "imagine" the future visual and geometric state of the world. This dense generation objective forces the model to internalize fine-grained knowledge about scene dynamics.

RGB Generation as Visual Supervision. The future RGB generation objective requires the model to predict the visual appearance of upcoming driving scenes, providing dense supervision over texture, color, object identity, and scene semantics. Each future RGB frame is tokenized via the shared MAGVIT-v2 quantizer, and the model learns to reconstruct randomly masked tokens through the mask token prediction objective. Formally, the RGB generation loss over all F future frames is:

$$\mathcal{L}_{\text{rgb}} = -\sum_{f=1}^{F}\sum_{j}\log p_\theta\big(u_{f,j}^{\text{rgb}} \mid \mathbf{u}_{f,*}^{\text{rgb}},\ \mathbf{v},\ \mathbf{e}_s,\ \mathbf{u}\big), \tag{2}$$

where $u_{f,j}^{\text{rgb}}$ is the j-th masked token of the f-th future RGB frame, and $\mathbf{u}_{f,*}^{\text{rgb}}$ denotes the corresponding masked sequence. By reconstructing future visual tokens, the model learns to capture how scene appearance evolves under the current driving context.

Depth Generation as Geometric Supervision. Complementary to RGB, the depth generation objective supervises the model on the 3D geometric structure of future scenes. Depth maps encode object distances, spatial layout, and surface orientation, which is critical for safe planning but entirely absent from text or trajectory supervision. The depth generation loss is defined analogously:

$$\mathcal{L}_{\text{depth}} = -\sum_{f=1}^{F}\sum_{j}\log p_\theta\big(u_{f,j}^{\text{dep}} \mid \mathbf{u}_{f,*}^{\text{dep}},\ \mathbf{u}_{f,*}^{\text{rgb}},\ \mathbf{v},\ \mathbf{e}_s,\ \mathbf{u}\big), \tag{3}$$

where $u_{f,j}^{\text{dep}}$ is the j-th masked token of the f-th future depth map. The overall mask token prediction (MTP) loss decomposes as:

$$\mathcal{L}_{\text{MTP}} = \mathcal{L}_{\text{rgb}} + \mathcal{L}_{\text{depth}}. \tag{4}$$

3.5 Intrinsic Reward from World Model for Exploration

Models trained purely via behavior cloning tend to narrowly replicate the expert's actions and struggle to generalize when the test-time distribution deviates from the training data. In the second stage of training, we leverage the world model learned in the first stage as a source of intrinsic reward signals that encourage the policy to explore novel yet safe driving behaviors beyond mere imitation. The core idea is that the world model's image prediction uncertainty provides a natural measure of trajectory novelty relative to the entire training distribution: high uncertainty indicates that the model has rarely encountered the visual consequences of a given action, signaling an out-of-distribution trajectory that, if safe, represents a valuable learning opportunity.

Uncertainty-Based Exploration Bonus. Given a candidate trajectory $\boldsymbol{\tau}_i$ sampled from the current VLA policy, we condition the world model on $\boldsymbol{\tau}_i$ and perform future image generation. For each predicted discrete image token, the model outputs a probability distribution over the MAGVIT-v2 codebook. We quantify the prediction uncertainty using the entropy of these token-level distributions, averaged across all generated future RGB and depth tokens:

$$\mathcal{H}(\boldsymbol{\tau}_i) = -\frac{1}{|\mathcal{M}|} \sum_{j \in \mathcal{M}} p_j \log p_j, \tag{5}$$

where $\mathcal{M}$ denotes the set of generated image token positions across all future RGB and depth frames, and p_j is the predicted probability of image token j. A high entropy indicates that the world model is uncertain about the visual consequences of trajectory $\boldsymbol{\tau}_i$, meaning the trajectory leads to scenarios under-represented in the training distribution. Conversely, low entropy indicates an in-distribution trajectory whose consequences the model has already learned to predict. We define the exploration bonus as the normalized entropy:

$$b_i = f(\mathcal{H}(\boldsymbol{\tau}_i)), \tag{6}$$

where $f(\mathcal{H}) = (\mathcal{H} - \mathcal{H}_{\min})/(\mathcal{H}_{\max} - \mathcal{H}_{\min})$ maps the raw entropy to a normalized bonus $b_i \in [0, 1]$.

Safety-Gated Reward. High uncertainty alone does not guarantee that exploration is beneficial, since a trajectory leading to a collision is novel but not useful. We therefore gate the exploration bonus using the PDMS [7], which evaluates trajectory quality on a $[0, 1]$ scale based on collision avoidance, comfort, and progress. The intrinsic reward for trajectory $\boldsymbol{\tau}_i$ is:

$$R_i = \begin{cases} \text{PDMS}_i + \lambda \cdot b_i, & \text{if PDMS}_i > \delta, \\ \text{PDMS}_i, & \text{otherwise}, \end{cases} \tag{7}$$

where δ is a safety threshold and λ controls the exploration strength. The gating mechanism ensures that only safe trajectories ($\text{PDMS}_i > \delta$) receive the exploration bonus. Unsafe or failed explorations are scored purely by their PDMS without additional encouragement. This design prioritizes trajectories that are simultaneously novel (high world model uncertainty) and good (high PDMS). Such trajectories represent out-of-distribution behaviors that are critical for improving generalization beyond expert demonstrations.

Policy Optimization via GRPO. We optimize the policy using GRPO [34]. At each iteration, we sample a group of G candidate trajectories $\{\boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_G\}$ from the current policy. Each trajectory and corresponding predicted image tokens are scored using the reward in Eq. (7), and rewards are normalized within the group to compute relative advantages:

$$\hat{A}_i = \frac{R_i - \text{mean}(\{R_1, \cdots, R_G\})}{\text{std}(\{R_1, \cdots, R_G\})}. \tag{8}$$

The policy is updated to increase the likelihood of trajectories with positive advantages and suppress those with negative advantages. This group-relative formulation naturally steers the policy toward trajectories that are both good and novel within each sampled group.

3.6 Training Strategy

As shown in Fig. 2, our training proceeds in two stages. The first stage consists of two phases: pre-training and supervised fine-tuning. During pre-training, ground-truth future actions are provided as input, and the model is trained solely with the image generation objective. This allows the model to adapt its visual generation capabilities to driving scenes. In the supervised fine-tuning phase, the model is trained to jointly predict both actions and future images. In the second stage, we further refine the policy using an intrinsic reward derived from the world model to encourage exploration.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our method on two widely used autonomous driving benchmarks: NAVSIM and nuScenes.

NAVSIM [2, 7] is a recently proposed non-reactive simulation benchmark built upon the OpenScene dataset. NAVSIM v1 evaluates planning quality using PDMS [7], and NAVSIM v2 adopts the Extended PDMS (EPDMS) [2]. Both metrics aggregate factors such as progress, time-to-collision, and comfort, achieving stronger correlation with closed-loop evaluations. We train on the `navtrain` split and report results on the `navtest` split. In NAVSIM, the model predicts future waypoints in the form of (x, y, θ) over a 4-second planning horizon.

nuScenes [1] consists of 1,000 driving scenes and has become a standard benchmark for open-loop planning evaluation. Following prior works [18, 19, 22], we adopt the standard train/val split and evaluate using the L2 displacement error and collision rate between predicted and ground-truth trajectories. On nuScenes, the model predicts future waypoints in the form of (x, y). The evaluation results on nuScenes are presented in the supplementary material.

Implementation Details. Our model is built upon Show-o [39] as the backbone architecture. In the first training stage, we first pre-train the model for 10 epochs by conditioning on ground-truth future actions and supervising only the image generation. We then perform supervised fine-tuning for 15 epochs, where the model jointly predicts future trajectories and generates future RGB and depth images. In the second stage, we apply GRPO-based post-training with LoRA [17] for 5 epochs to further refine the policy using the exploration-aware reward described in Sec. 3.5. All experiments are conducted on 4×H200 GPUs. We obtain depth maps from a pre-trained monocular depth estimation model [46]. Detailed hyperparameters are provided in the supplementary material.

Table 1: Comparison on NAVSIM v1 with closed-loop metrics. The best performance is marked in **bold**, and the second best is underlined. Abbreviations: no at-fault collision (NC), drivable area compliance (DAC), ego progress (EP), time to collision (TTC), comfort (Comf.). SC: single-view camera; MC: multi-view camera; L: LiDAR; †: best-of-N (N=6) strategy [56].

Model	Input	NC ↑	DAC ↑	EP ↑	TTC ↑	Comf. ↑	PDMS ↑
Ego Status MLP	-	93.1	78.3	63.2	84.0	99.9	66.4
TransFuser [5]	-	97.8	92.6	78.9	92.9	99.9	83.9
DRAMA [48]	MC+L	98.2	95.2	81.3	94.2	**100.0**	86.9
Hydra-MDP [30]	MC+L	99.1	98.3	85.2	96.6	**100.0**	91.3
Centaur [35]	MC+L	99.2	98.7	86.0	97.2	99.9	92.1
DriveSuprim [45]	MC+L	98.6	98.6	**91.3**	95.5	**100.0**	93.5
DrivingGPT [4]	SC	98.9	90.7	79.9	94.9	95.6	82.4
FSDrive [50]	SC	98.2	93.8	80.1	93.3	99.9	85.1
PWM [52]	SC	98.9	95.8	81.5	95.9	**100.0**	88.1
AutoVLA [56]	MC	98.4	95.6	81.9	98.0	99.9	89.1
AutoVLA† [56]	MC	99.1	97.1	87.6	97.1	99.9	92.1
DriveVLA-W0 [27]	SC	98.7	**99.1**	87.6	97.1	**100.0**	90.2
DriveVLA-W0† [27]	SC	**99.9**	97.4	88.3	97.0	99.9	93.0
ExploreVLA	SC	98.8	98.4	83.5	96.5	99.9	90.4
ExploreVLA†	SC	99.4	98.9	88.3	**98.3**	99.7	**93.7**

4.2 Main Results

Results on NAVSIM v1. Tab. 1 presents the comparison on the NAVSIM v1 benchmark. Our ExploreVLA achieves the highest PDMS of 93.7 with the best-of-N strategy, outperforming all prior approaches. Notably, ExploreVLA uses only a single-view camera, yet surpasses multi-sensor methods such as DriveSuprim and Centaur. Even without the best-of-N strategy, ExploreVLA attains a PDMS of 90.4, which is competitive with multi-view methods like AutoVLA. Among the sub-metrics, ExploreVLA† achieves the best TTC and the second-best scores on NC, DAC, and EP, demonstrating that our world-model-based exploration reward helps the model internalize diverse and robust driving behaviors beyond what pure imitation learning can provide.

Results on NAVSIM v2. Tab. 2 further validates our approach on NAVSIM v2, which introduces extended closed-loop metrics including driving direction compliance (DDC), traffic light compliance (TLC), lane keeping (LK), history comfort (HC), and extended comfort (EC). ExploreVLA achieves the highest EPDMS of 88.8, surpassing the previous best result of 86.1 by DriveVLA-W0 by 2.7 points. Our method obtains the best scores on six out of nine individual metrics, while remaining highly competitive on the rest.

Table 2: Comparison on NAVSIM v2 with extended closed-loop metrics. The best performance is marked in **bold**, and the second best is underlined.

Model	NC ↑	DAC ↑	DDC ↑	TLC ↑	EP ↑	TTC ↑	LK ↑	HC ↑	EC ↑	EPDMS ↑
Ego Status	93.1	77.9	92.7	99.6	86.0	91.5	89.4	**98.3**	85.4	64.0
TransFuser [5]	96.9	89.9	97.8	99.7	87.1	95.4	92.7	**98.3**	87.2	76.7
Hydra-MDP++ [26]	97.2	97.5	99.4	99.6	83.1	96.5	94.4	98.2	70.9	81.4
DriveSuprem [45]	97.5	96.5	99.4	99.6	88.4	96.6	95.5	**98.3**	77.0	83.1
ARTEMIS [9]	98.3	95.1	98.6	**99.8**	81.5	97.4	96.5	**98.3**	-	83.1
DiffusionDrive [31]	98.2	95.9	99.4	**99.8**	87.5	97.3	96.8	**98.3**	87.7	84.5
DiffusionDriveV2 [57]	97.7	96.6	99.2	**99.8**	**88.9**	97.2	96.0	97.8	**91.0**	85.5
DriveVLA-W0 [27]	98.5	**99.1**	98.0	99.7	86.4	98.1	93.2	97.9	58.9	86.1
ExploreVLA	**98.8**	96.2	**99.6**	**99.8**	87.1	**98.2**	**97.8**	**98.3**	86.8	**88.8**

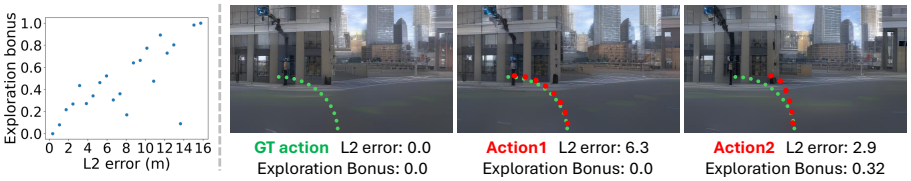

Fig. 3: Analysis of the exploration bonus. Left: the exploration bonus is positively correlated with L2 error to the ground-truth trajectory. Right: our exploration bonus can properly measure the trajectory novelty that L2 error fails.

4.3 Analysis of Intrinsic Reward Modeling

Fig. 3 provides an analysis of our intrinsic reward modeling mechanism. The left shows a generally positive correlation between the exploration bonus and L2 error with respect to the ground-truth trajectory: as the sampled trajectory deviates further from the expert action, the world model's prediction uncertainty tends to increase, resulting in higher exploration bonuses. However, L2 distance is not always an unreliable measure of novelty, and our uncertainty-based bonus reflects it more faithfully. The right of Fig. 3 shows a representative failure case: a trajectory that closely follows the expert's direction may incur a large L2 error due to positional shift, while a trajectory that takes a fundamentally different route may have a smaller L2 error. Additionally, we randomly select 1,000 straight-driving scenes and apply two types of perturbation to the expert trajectory: (1) speed variation, which changes vehicle speed while following nearly identical paths, and (2) direction variation, which changes the driving direction and therefore represents genuinely different driving behaviors. As shown in Tab. 3, the speed perturbation produces a substantially larger L2 error despite exhibiting little behavioral novelty, whereas the direction perturbation yields a smaller L2 error while receiving a significantly higher exploration bonus. These results indicate that the proposed uncertainty-based reward better captures behavioral novelty than trajectory distance.

Table 3: L2 error vs. exploration bonus on 1,000 randomly selected straight-driving scenes under two perturbation types. L2 error misjudges novelty in both cases, whereas our exploration bonus correctly assigns low novelty to speed changes (similar path) and high novelty to direction changes (different path).

	Type 1 (speed Δ, similar path)	Type 2 (direction Δ, different path)
L2 (m)	5.82 ± 1.74	2.15 ± 0.93
Exploration bonus	0.04 ± 0.03	0.29 ± 0.12

Table 4: Ablation study on dense visual supervision. We evaluate the effect of auxiliary RGB and depth image generation during Stage 1 imitation learning on the NAVSIM v1 `navtest` split.

RGB Img. Gen.	Depth Img. Gen.	NC ↑	DAC ↑	EP ↑	TTC ↑	Comf. ↑	PDMS ↑
✗	✗	98.6	94.4	80.1	94.8	**100.0**	86.2
✓	✗	**98.7**	96.0	81.5	95.4	99.9	87.9
✗	✓	**98.7**	95.8	81.4	95.6	99.9	87.8
✓	✓	**98.7**	**96.3**	**82.3**	**95.7**	99.9	**88.5**

4.4 Ablation Study

Effect of Dense Visual Supervision. Tab. 4 examines the contribution of RGB and depth image generation as auxiliary supervision during Stage 1 training. The trajectory-only baseline without any image generation achieves the lowest PDMS. Adding either RGB or depth generation alone yields comparable improvements, which confirms that both modalities provide meaningful dense supervisory signals for learning better scene representations. Combining both RGB and depth generation further improves PDMS to 88.5. This indicates that RGB and depth capture complementary information (visual appearance and geometric structure) and their joint supervision leads to a more comprehensive understanding of driving scenes.

Effect of Reward Design. Tab. 5 isolates the contribution of each reward component during Stage 2 RL. Starting from the Stage 1 model (PDMS 88.50), applying only the PDMS reward brings a substantial improvement to 90.19, demonstrating the effectiveness of GRPO-based post-training. Using the image-based exploration reward alone yields only a marginal gain (88.53), which is expected since the image reward serves as an exploration bonus rather than a direct driving quality signal. Without the safety gate provided by PDMS thresholding, the exploration signal alone cannot effectively guide policy improvement. When combining both rewards, the model achieves the best PDMS of 90.36. This confirms that the image-based exploration reward provides a complementary learning signal that encourages the policy to discover diverse driving strategies beyond what the PDMS reward alone can incentivize.

Table 5: Ablation study on reward components. We evaluate the contribution of each reward signal during Stage 2 reinforcement learning on the NAVSIM v1 `navtest` split. The first row (no reward) corresponds to the Stage 1 baseline.

PDMS Reward	Image Reward	NC ↑	DAC ↑	EP ↑	TTC ↑	Comf. ↑	PDMS ↑
✗	✗	98.74	96.25	82.27	95.67	**99.97**	88.50
✓	✗	**98.82**	97.78	83.44	96.50	99.95	90.19
✗	✓	98.76	96.30	82.32	95.65	**99.97**	88.53
✓	✓	98.81	**97.88**	**83.53**	**96.66**	99.95	**90.36**

Table 6: Ablation on the safety threshold δ. We evaluate the effect of different safety thresholds in Eq. (7) during Stage 2 reinforcement learning on the NAVSIM v1 `navtest` split.

δ	NC ↑	DAC ↑	EP ↑	TTC ↑	Comf. ↑	PDMS ↑
0.0	**98.85**	97.82	83.48	96.53	99.92	90.25
0.3	98.83	97.78	83.47	96.49	99.91	90.21
0.6	98.81	97.79	83.44	96.54	99.93	90.22
0.9	98.81	**97.88**	**83.53**	**96.66**	**99.95**	**90.36**

Effect of the Safety Threshold. The safety threshold δ in Eq. (7) controls which trajectories are eligible for the exploration bonus. We ablate $\delta \in \{0, 0.3, 0.6, 0.9\}$ in Tab. 6. The proposed exploration reward consistently improves over the PDMS-only baseline (second row in Tab. 5) across all threshold choices. This result indicates that the effectiveness of our method is not sensitive to the exact value of δ. Among all settings, $\delta = 0.9$ achieves the best performance. The similar performance observed for $\delta \in \{0, 0.3, 0.6\}$ can be explained by the Stage 1 policy distribution. Approximately 99% of the sampled trajectories already achieve PDMS scores above 0.6, and thus making these thresholds insufficient for distinguishing safe trajectories from risky ones. When δ is increased to 0.9, the safety gate begins to effectively separate safe-and-novel trajectories from unsafe exploratory behaviors, leading to the highest PDMS.

Qualitative Analysis. Fig. 4 visualizes planned trajectories after Stage 1 and Stage 2 in three representative scenarios. In each case, the Stage 1 model exhibits a safety-critical failure (*i.e.*, colliding with a vehicle, passing dangerously close to pedestrians, or running a stop sign). After Stage 2 reinforcement learning, the model successfully corrects these behaviors. These results demonstrate that our world-model-based RL not only improves aggregate metrics but also rectifies safety-critical failures that pure imitation learning struggles to resolve from expert demonstrations alone.

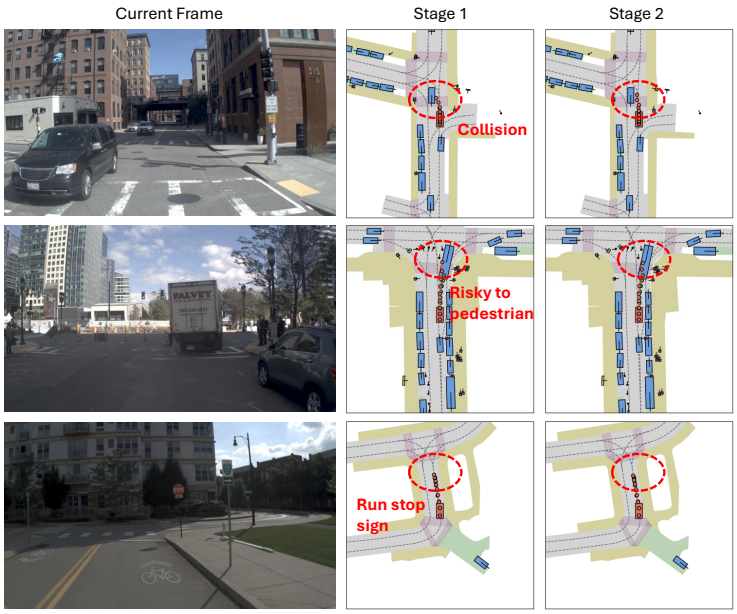

Fig. 4: Qualitative comparison of planned trajectories before and after RL post-training. We visualize three challenging driving scenarios in bird's-eye view. The Stage 1 model exhibits safety-critical failures. After Stage 2 RL post-training, the model produces safer and more compliant trajectories. green: GT, orange: prediction.

5 Conclusion

We presented ExploreVLA, a unified framework that addresses the lack of exploration and sparse supervision in VLA-based autonomous driving. By jointly predicting future trajectories, RGB images, and depth maps, our approach provides dense world modeling supervision that enriches scene representations. We further leverage the world model's prediction uncertainty as an intrinsic novelty measure, combined with a safety-gated reward and GRPO, to guide the policy toward diverse yet safe driving strategies beyond expert imitation. Extensive experiments on the NAVSIM and nuScenes benchmarks validate the effectiveness of our approach, achieving state-of-the-art performance on NAVSIM.

Limitations and Future Work. Our framework currently uses a single front-view camera; extending to multi-view inputs could further broaden spatial coverage and planning robustness. Moreover, our reinforcement learning stage relies on offline/open-loop post-training, which may bias the policy toward the training distribution and provide limited opportunities to learn recovery behaviors from self-induced states. Future work will investigate integrating our uncertainty-guided exploration framework with large-scale closed-loop reinforcement learning and evaluation.

References

1. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
2. Cao, W., Hallgarten, M., Li, T., Dauner, D., Gu, X., Wang, C., Miron, Y., Aiello, M., Li, H., Gilitschenski, I., et al.: Pseudo-simulation for autonomous driving. arXiv preprint arXiv:2506.04218 (2025)
3. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
4. Chen, Y., Wang, Y., Zhang, Z.: Drivinggpt: Unifying driving world modeling and planning with multi-modal autoregressive transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26890–26900 (2025)
5. Chitta, K., Prakash, A., Jaeger, B., Yu, Z., Renz, K., Geiger, A.: Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(11), 12878–12895 (2022)
6. Codevilla, F., Santana, E., López, A.M., Gaidon, A.: Exploring the limitations of behavior cloning for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9329–9338 (2019)
7. Dauner, D., Hallgarten, M., Li, T., Weng, X., Huang, Z., Yang, Z., Li, H., Gilitschenski, I., Ivanovic, B., Pavone, M., et al.: Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. Advances in Neural Information Processing Systems **37**, 28706–28719 (2024)
8. Ding, J., Zhang, Y., Shang, Y., Zhang, Y., Zong, Z., Feng, J., Yuan, Y., Su, H., Li, N., Sukiennik, N., et al.: Understanding world or predicting future? a comprehensive survey of world models. ACM Computing Surveys **58**(3), 1–38 (2025)
9. Feng, R., Xi, N., Chu, D., Wang, R., Deng, Z., Wang, A., Lu, L., Wang, J., Huang, Y.: Artemis: Autoregressive end-to-end trajectory planning with mixture of experts for autonomous driving. IEEE Robotics and Automation Letters **11**(1), 226–233 (2025)
10. Feng, T., Wang, W., Yang, Y.: A survey of world models for autonomous driving. arXiv preprint arXiv:2501.11260 (2025)
11. Fu, H., Zhang, D., Zhao, Z., Cui, J., Liang, D., Zhang, C., Zhang, D., Xie, H., Wang, B., Bai, X.: Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24823–24834 (2025)
12. Gao, S., Yang, J., Chen, L., Chitta, K., Qiu, Y., Geiger, A., Zhang, J., Li, H.: Vista: A generalizable driving world model with high fidelity and versatile controllability. Advances in Neural Information Processing Systems **37**, 91560–91596 (2024)
13. Guan, Y., Liao, H., Li, Z., Hu, J., Yuan, R., Zhang, G., Xu, C.: World models for autonomous driving: An initial survey. IEEE Transactions on Intelligent Vehicles (2024)
14. Guo, Y., Zhang, J., Chen, X., Ji, X., Wang, Y.J., Hu, Y., Chen, J.: Improving vision-language-action model with online reinforcement learning. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 15665–15672. IEEE (2025)
15. Hassan, M., Stapf, S., Rahimi, A., Rezende, P., Haghighi, Y., Brüggemann, D., Katircioglu, I., Zhang, L., Chen, X., Saha, S., et al.: Gem: A generalizable ego-vision

multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22404–22415 (2025)
16. Hu, A., Russell, L., Yeo, H., Murez, Z., Fedoseev, G., Kendall, A., Shotton, J., Corrado, G.: Gaia-1: A generative world model for autonomous driving. arXiv preprint arXiv:2309.17080 (2023)
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
18. Hu, S., Chen, L., Wu, P., Li, H., Yan, J., Tao, D.: St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In: European Conference on Computer Vision. pp. 533–549. Springer (2022)
19. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., et al.: Planning-oriented autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17853–17862 (2023)
20. Hwang, J.J., Xu, R., Lin, H., Hung, W.C., Ji, J., Choi, K., Huang, D., He, T., Covington, P., Sapp, B., et al.: Emma: End-to-end multimodal model for autonomous driving. arXiv preprint arXiv:2410.23262 (2024)
21. Jiang, A., Gao, Y., Sun, Z., Wang, Y., Wang, J., Chai, J., Cao, Q., Heng, Y., Jiang, H., Dong, Y., et al.: Diffvla: Vision-language guided diffusion planning for autonomous driving. arXiv preprint arXiv:2505.19381 (2025)
22. Jiang, B., Chen, S., Xu, Q., Liao, B., Chen, J., Zhou, H., Zhang, Q., Liu, W., Huang, C., Wang, X.: Vad: Vectorized scene representation for efficient autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8340–8350 (2023)
23. Jiang, S., Huang, Z., Qian, K., Luo, Z., Zhu, T., Zhong, Y., Tang, Y., Kong, M., Wang, Y., Jiao, S., et al.: A survey on vision-language-action models for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4524–4536 (2025)
24. Li, B., Guo, J., Liu, H., Zou, Y., Ding, Y., Chen, X., Zhu, H., Tan, F., Zhang, C., Wang, T., et al.: Uniscene: Unified occupancy-centric driving scene generation. In: Proceedings of the computer vision and pattern recognition conference. pp. 11971–11981 (2025)
25. Li, B., Ma, Z., Du, D., Peng, B., Liang, Z., Liu, Z., Ma, C., Jin, Y., Zhao, H., Zeng, W., et al.: Omninwm: Omniscient driving navigation world models. arXiv preprint arXiv:2510.18313 (2025)
26. Li, K., Li, Z., Lan, S., Xie, Y., Zhang, Z., Liu, J., Wu, Z., Yu, Z., Alvarez, J.M.: Hydra-mdp++: Advancing end-to-end driving via expert-guided hydra-distillation. arXiv preprint arXiv:2503.12820 (2025)
27. Li, Y., Shang, S., Liu, W., Zhan, B., Wang, H., Wang, Y., Chen, Y., Wang, X., An, Y., Tang, C., et al.: Drivevla-w0: World models amplify data scaling law in autonomous driving. arXiv preprint arXiv:2510.12796 (2025)
28. Li, Y., Wang, Y., Liu, Y., He, J., Fan, L., Zhang, Z.: End-to-end driving with online trajectory evaluation via bev world model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27137–27146 (2025)
29. Li, Y., Xiong, K., Guo, X., Li, F., Yan, S., Xu, G., Zhou, L., Chen, L., Sun, H., Wang, B., et al.: Recogdrive: A reinforced cognitive framework for end-to-end autonomous driving. arXiv preprint arXiv:2506.08052 (2025)
30. Li, Z., Li, K., Wang, S., Lan, S., Yu, Z., Ji, Y., Li, Z., Zhu, Z., Kautz, J., Wu, Z., et al.: Hydra-mdp: End-to-end multimodal planning with multi-target hydra-distillation. arXiv preprint arXiv:2406.06978 (2024)

31. Liao, B., Chen, S., Yin, H., Jiang, B., Wang, C., Yan, S., Zhang, X., Li, X., Zhang, Y., Zhang, Q., et al.: Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12037–12047 (2025)
32. Min, C., Zhao, D., Xiao, L., Nie, Y., Dai, B.: Uniworld: Autonomous driving pretraining via world models. arXiv preprint arXiv:2308.07234 (2023)
33. Ross, S., Gordon, G., Bagnell, D.: 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. vol. 15, pp. 627–635. PMLR (2011)
34. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
35. Sima, C., Chitta, K., Yu, Z., Lan, S., Luo, P., Geiger, A., Li, H., Alvarez, J.M.: Centaur: Robust end-to-end autonomous driving with test-time training. arXiv preprint arXiv:2503.11650 (2025)
36. Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-drive world models for autonomous driving. In: European conference on computer vision. pp. 55–72. Springer (2024)
37. Wang, X., Cui, Y., Wang, J., Zhang, F., Wang, Y., Zhang, X., Luo, Z., Sun, Q., Li, Z., Wang, Y., et al.: Multimodal learning with next-token prediction for large multimodal models. Nature pp. 1–7 (2026)
38. Wang, Y., Luo, W., Bai, J., Cao, Y., Che, T., Chen, K., Chen, Y., Diamond, J., Ding, Y., Ding, W., et al.: Alpamayo-r1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail. arXiv preprint arXiv:2511.00088 (2025)
39. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. arXiv preprint arXiv:2408.12528 (2024)
40. Xiong, Z., Ye, X., Yaman, B., Cheng, S., Lu, Y., Luo, J., Jacobs, N., Ren, L.: Unidrive-wm: Unified understanding, planning and generation world model for autonomous driving. arXiv preprint arXiv:2601.04453 (2026)
41. Xu, Z., Zhang, Y., Xie, E., Zhao, Z., Guo, Y., Wong, K.Y.K., Li, Z., Zhao, H.: Drivegpt4: Interpretable end-to-end autonomous driving via large language model. IEEE Robotics and Automation Letters **9**(10), 8186–8193 (2024)
42. Yan, T., Tang, T., Gui, X., Li, Y., Zhesng, J., Huang, W., Kong, L., Han, W., Zhou, X., Zhang, X., et al.: Ad-r1: Closed-loop reinforcement learning for end-to-end autonomous driving with impartial world models. arXiv preprint arXiv:2511.20325 (2025)
43. Yang, J., Chitta, K., Gao, S., Chen, L., Shao, Y., Jia, X., Li, H., Geiger, A., Yue, X., Chen, L.: Resim: Reliable world simulation for autonomous driving. arXiv preprint arXiv:2506.09981 (2025)
44. Yang, J., Gao, S., Qiu, Y., Chen, L., Li, T., Dai, B., Chitta, K., Wu, P., Zeng, J., Luo, P., et al.: Generalized predictive model for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14662–14672 (2024)
45. Yao, W., Li, Z., Lan, S., Wang, Z., Sun, X., Alvarez, J.M., Wu, Z.: Drivesuprim: Towards precise trajectory selection for end-to-end planning. arXiv preprint arXiv:2506.06659 (2025)

46. Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., Chen, X., Shen, C.: Metric3d: Towards zero-shot metric 3d prediction from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9043–9053 (2023)
47. Yu, L., Lezama, J., Gundavarapu, N.B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Birodkar, V., Gupta, A., Gu, X., et al.: Language model beats diffusion–tokenizer is key to visual generation. arXiv preprint arXiv:2310.05737 (2023)
48. Yuan, C., Zhang, Z., Sun, J., Sun, S., Huang, Z., Lee, C.D.W., Li, D., Han, Y., Wong, A., Tee, K.P., et al.: Drama: An efficient end-to-end motion planner for autonomous driving with mamba. arXiv preprint arXiv:2408.03601 (2024)
49. Yuan, J., Sun, S., Omeiza, D., Zhao, B., Newman, P., Kunze, L., Gadd, M.: Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model. arXiv preprint arXiv:2402.10828 (2024)
50. Zeng, S., Chang, X., Xie, M., Liu, X., Bai, Y., Pan, Z., Xu, M., Wei, X., Guo, N.: Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. Advances in Neural Information Processing Systems (2025)
51. Zhang, K., Tang, Z., Hu, X., Pan, X., Guo, X., Liu, Y., Huang, J., Yuan, L., Zhang, Q., Long, X.X., et al.: Epona: Autoregressive diffusion world model for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27220–27230 (2025)
52. Zhao, Z., Fu, T., Wang, Y., Wang, L., Lu, H.: From forecasting to planning: Policy world model for collaborative state-action prediction. In: Advances in Neural Information Processing Systems (2025)
53. Zheng, W., Chen, W., Huang, Y., Zhang, B., Duan, Y., Lu, J.: Occworld: Learning a 3d occupancy world model for autonomous driving. In: European conference on computer vision. pp. 55–72. Springer (2024)
54. Zheng, Y., Yang, P., Xing, Z., Zhang, Q., Zheng, Y., Gao, Y., Li, P., Zhang, T., Xia, Z., Jia, P., et al.: World4drive: End-to-end autonomous driving via intention-aware physical latent world model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 28632–28642 (2025)
55. Zhou, X., Han, X., Yang, F., Ma, Y., Knoll, A.C.: Opendrivevla: Towards end-to-end autonomous driving with large vision language action model. arXiv preprint arXiv:2503.23463 (2025)
56. Zhou, Z., Cai, T., Zhao, S.Z., Zhang, Y., Huang, Z., Zhou, B., Ma, J.: 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 (2025)
57. Zou, J., Chen, S., Liao, B., Zheng, Z., Song, Y., Zhang, L., Zhang, Q., Liu, W., Wang, X.: Diffusiondrivev2: Reinforcement learning-constrained truncated diffusion modeling in end-to-end autonomous driving. arXiv preprint arXiv:2512.07745 (2025)