# HyWorldVLA: A Vision-Language-Action Model with Hybrid World Modeling for Autonomous Driving

**Anonymous authors**
Paper under double-blind review

## Abstract

Vision-Language-Action (VLA) models augmented with world modeling represent a promising paradigm for end-to-end autonomous driving. While pixel-level future prediction enables fine-grained spatiotemporal reasoning, it compromises robustness in noisy driving scenarios. Conversely, latent-based world models alleviate this sensitivity but often incur limited interpretability and representational degradation due to absent pixel-level grounding. To reconcile this trade-off, we propose HyWorldVLA, a hybrid world-VLA framework that unifies pixel-level supervision and latent representation learning. In the pre-training stage, HyWorldVLA predicts video latents encoded by a pre-trained video VAE, while simultaneously reconstructing video frames to provide precise pixel-level grounding. During the subsequent co-fine-tuning phase, the model exclusively predicts latent features, which are fed into an action expert to generate trajectories. Extensive experiments on NAVSIM v1 and v2 benchmarks demonstrate that HyWorldVLA outperforms both pixel-based and latent-based world model baselines. Notably, we present the first comprehensive qualitative and quantitative analysis of world model noise robustness in autonomous driving, establishing a new benchmark for evaluating future architectures.

## 1 Introduction

End-to-end autonomous driving has emerged as a dominant research direction for building intelligent vehicle systems, aiming to unify perception, reasoning, and control within a single holistic pipeline. As a core driving-centric paradigm, Vision-Language-Action (VLA) models (Wang et al., 2023; Hwang et al., 2024; Li et al., 2025e; Lu et al., 2026a) integrate visual scene perception, language semantic understanding, and low-level action prediction, effectively leveraging large-scale driving datasets and generalized vision-language priors to learn robust driving behaviors. Compared with traditional modular autonomous driving systems that suffer from cascading errors and limited generalization on long-tailed and interactive traffic scenarios, VLA-based frameworks deliver more adaptive and generalizable decision-making capabilities, showing great potential for real-world vehicle deployment.

To further strengthen the spatiotemporal reasoning and scenario generalization of VLA models, recent advances augment standard driving VLA pipelines with world modeling capabilities, which learn to predict future scene states for forward-looking causal reasoning (Jia et al., 2023; Wang et al., 2024a; Chen et al., 2025). Existing world modeling paradigms for driving VLAs fall into two dominant categories with inherent trade-offs. Pixel-based world models enforce supervision via full future frame reconstruction, providing dense geometric and physical constraints that yield strong generalization on occluded, interactive, and long-tailed driving scenarios, as shown in Figure 1 (a). However, pixel-level reconstruction is inherently sensitive to environmental variations ranging from rain and fog to fluctuating illumination, which fundamentally constrains the deployment robustness of pixel-based world models under complex driving conditions. In contrast, latent-based world VLAs improve scene noise robustness by forecasting future scene states purely in compact feature space, as shown in Figure 1 (b). Nevertheless, the absence of explicit pixel-level grounding results in representation degeneration and limited interpretability, which severely deteriorates model performance in complex real-world driving scenarios (Tu et al., 2025).



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

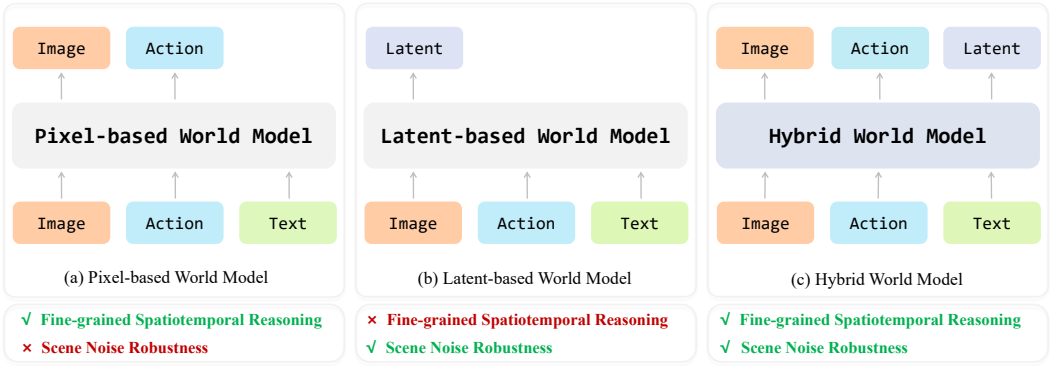



Figure 1: **Comparison of world model paradigms.** (a) Pixel-based World Model: iteratively predicts actions and images, prone to scene noise; (b) Latent-based World Model: predicts latent states, weak in spatiotemporal dynamics modeling; (c) Hybrid World Model: our method simultaneously models iterative action-image prediction and future latent state prediction.

To resolve this fundamental dilemma in the world modeling domain, we present HyWorldVLA, a novel hybrid world-VLA framework that unifies the strengths of pixel-grounded supervision and latent feature prediction to deliver both accuracy and robustness for end-to-end autonomous driving. Unlike existing purely pixel-based or latent-based world models, as shown in Figure 1 (c), HyWorldVLA employs a dual-supervision paradigm in the pre-training stage, jointly optimizing latent prediction and raw pixel frame reconstruction, where the latent features are extracted from language-guided VAE encoder. During the subsequent co-fine-tuning phase, the model exclusively predicts latent features, which are fed into an action expert to generate driving trajectories. Under this design, pixel-level reconstruction serves as a rigorous structural regularizer during pre-training to refine latent embedding learning and prevent representation collapse. For the co-fine-tuning stage, action generation is conditioned on latent feature modeling that retains compact temporal representations, bolstering the scene noise robustness of the generated driving trajectory. Extensive experimental results demonstrate that HyWorldVLA successfully combines the high performance brought by pixel-based world models' fine-grained spatiotemporal dynamic modeling with the scene noise robustness intrinsic to latent-based world models. Our contributions are summarized as follows:

- We propose HyWorldVLA, which establishes a hybrid world modeling paradigm that unifies the complementary characteristics of world action models and latent world models into a single framework.
- We introduce language guidance and latent conditioning mechanisms to effectively steer the model toward semantic information relevant to action generation.
- Extensive experiments on the NAVSIM benchmark demonstrate that our method achieves state-of-the-art performance, validating the effectiveness of our hybrid world modeling.

## 2 Related Work

**Vision-Language-Action Models in Driving.** With remarkable multimodal scene understanding and reasoning capabilities, Vision-Language Models (VLMs) are being extensively integrated into autonomous driving systems, promoting the development of end-to-end Vision-Language-Action (VLA) architectures (Hu et al., 2025). Early explorations including DriveMLM (Wang et al., 2023) and DriveGPT4 (Xu et al., 2024) directly leverage VLMs' inherent text-generation capabilities to produce textual driving actions, thus fully utilizing the world knowledge from pre-trained VLMs (Hwang et al., 2024; Sima et al., 2024; Jiang et al., 2025). However, the fundamental gap between discrete language tokens and continuous control spaces introduces precision limits and breaks the end-to-end differentiability. Consequently, recent works augment VLM backbones with mechanisms that produce direct numeric outputs. ReCogDrive (Li et al., 2025e) integrates an autoregressive model with a diffusion planner through a hierarchical cognitive pipeline, while AutoVLA (Zhou et al., 2025b) unifies semantic reasoning and trajectory planning in a single autoregressive model with dual thinking modes. ORION (Fu et al., 2025) and OpenDriveVLA (Zhou et al., 2025a) fur-

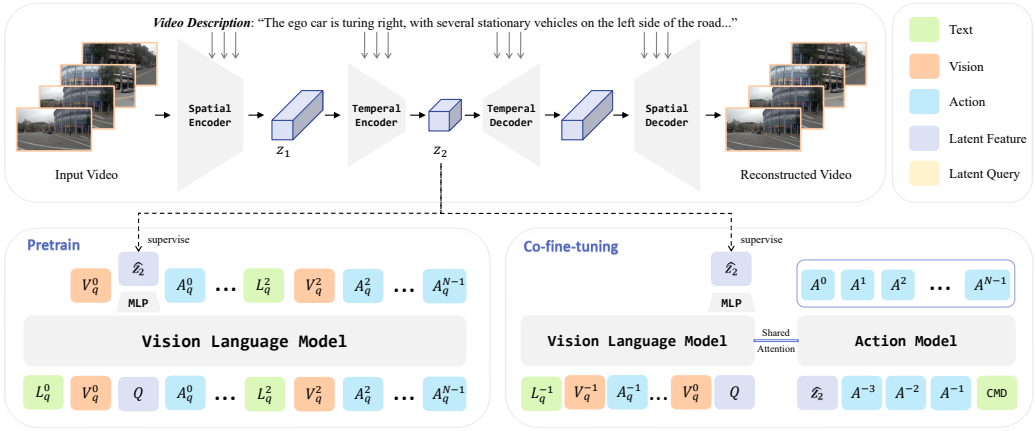



Figure 2: **Overview of HyWorldVLA.** HyWorldVLA consists of three training stages. First, a video VAE model is trained under text guidance to compress and reconstruct video frames, thereby providing ground-truth targets for subsequent training. In the pre-training stage, sequences composed of text, images, and actions are used to train the world model for iterative prediction, where a learnable latent query $Q$ is introduced to aggregate information and predict future latents. In the co-fine-tuning stage, the historical sequences together with $Q$ are fed into the VLM to obtain global context and predicted latents, which are then passed through a joint attention mechanism to the action expert for action generation.

ther bridge the semantic-action gap by aggregating long-term context and performing hierarchical vision-language alignment.

**World Models for Autonomous Driving.** World models have been applied to autonomous driving (Zheng et al., 2024; Hu et al., 2023a; Jia et al., 2023; Wang et al., 2024a; Li et al., 2025d; Chen et al., 2025; Zhang et al., 2025) due to their ability to anticipate future states and comprehend the temporal dynamics of the world. FSDrive (Zeng et al., 2025) generates unified future frames with explicit physical priors as visual spatiotemporal chain-of-thought. DriveVLA-W0 (Li et al., 2025c) addresses the "supervision deficit" in VLAs by predicting future images as dense self-supervised signals, instantiated via both autoregressive and diffusion world models. OneVL (Lu et al., 2026a) routes latent reasoning through compact tokens supervised by dual auxiliary decoders, matching answer-only latency. Latent-CoT-Drive (Tan et al., 2025) represents chain-of-thought reasoning in an action-aligned latent space using action-proposal and world model tokens. Pixel-level prediction is inherently susceptible to scene noise, whereas latent-level prediction struggles to capture the underlying causal relationships of the physical world. To address these complementary limitations, we propose a novel framework that bridges pixel-level and latent-level prediction.

# 3 Methodology

## 3.1 Problem Definition and Overall Framework

HyWorldVLA is formulated to predict future ego-vehicle waypoints conditioned on multi-modal observations. Specifically, at each time step $t$, the model receives a sequence of front-view camera images $\mathbf{V}_{t-H:t} = \{v_{t-H}, \ldots, v_t\}$, including the current and $H$ preceding frames, and historical vehicle waypoints $\mathbf{W}_{t-H:t} = \{w_{t-H}, \ldots, w_t\}$, where $w_t = (x_t, y_t, \theta_t)$ represents the position and heading angle at step $t$. Additionally, historical navigation commands $\mathbf{L}_{t-H:t}$ (e.g., *Go Straight* and *Turn Right*) are used to specify intended directions explicitly. Then based on these inputs, the model outputs a sequence of future waypoints $\mathbf{W}_{t+1:t+T}$ over a horizon of $T$ time steps.

As shown in Figure 2, the overall framework of HyWorldVLA consists of three main components and three training stages. *(1) Video Encoder:* Sec. 3.2 elaborates on a video encoder that compresses a sequence of video frames into a latent feature, providing a compact representation of future world dynamics; *(2) VLM Backbone:* Then the latent feature and video frames are jointly predicted by a VLM backbone, which is presented in Sec. 3.3; *(3) Action Expert:* Finally, in Sec. 3.4 we introduce the co-fine-tuning of the VLM model and a query-based action model, which predicts trajectories

conditioning on the hidden states from each layer of the VLM backbone, the predicted future visual latents, and the ego-vehicle's historical motion.

### 3.2 Language-guided Latent Feature Learning

To encode future frames into compact latent features for VLM supervision, we adopt a video variational autoencoder (Xing et al., 2024). As shown in Figure 2, taking a video segment $\mathbf{v} \in \mathbb{R}^{C\times T\times H\times W}$ as input, the model first employs a spatial encoder to produce spatial latent $\mathbf{z}_1 \in \mathbb{R}^{c\times T\times \frac{H}{8}\times \frac{W}{8}}$. Built upon a Stable Diffusion image VAE (Rombach et al., 2022), this spatial encoder stacks 3D convolutions to perform $8\times$ spatial downsampling while retaining full temporal length. Afterwards, a lightweight temporal autoencoder takes $\mathbf{z}_1$ as input and stacks 3D ResNet blocks to implement $4\times$ temporal compression, generating latent $\mathbf{z}_2 \in \mathbb{R}^{c'\times \frac{T}{4}\times \frac{H}{8}\times \frac{W}{8}}$ that eliminates redundant inter-frame static information. Then during reconstruction, the full-resolution video $\hat{\mathbf{v}}$ is recovered through a symmetric two-stage decoding pipeline.

To further enhance reconstruction quality, we embed multi-layer text cross-attention inside each block to introduce textual guidance during feature encoding. We split visual features into patch tokens, take visual patches as query and value, and adopt Flan-T5 (Chung et al., 2024) text embeddings as keys to compute cross-attention. The attention output is added back to original visual features via residual connection. This textual semantic prior effectively suppresses motion ghosting, edge blurring and temporal flickering, greatly improving reconstruction fidelity for complex traffic scenes.

The VAE is optimized via a multi-component loss that balances reconstruction loss $\mathcal{L}_{\text{rec}}$, adversarial loss $\mathcal{L}_{\text{GAN}}$, and KL-divergence regularization loss $\mathcal{L}_{\text{KL}}$:

$$\mathcal{L}_{\text{vae}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{GAN}}\mathcal{L}_{\text{GAN}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}}. \tag{1}$$

Through hierarchical spatial-temporal compression and text cross-modal injection, the video VAE generates a compact latent representation, providing effective supervision for downstream VLA pre-training and co-fine-tuning.

### 3.3 Pre-training

The pre-training stage aims to equip the backbone with the ability to predict the world evolution both on the image-level and latent-level.

**Vision Modeling.** We follow VQGAN (Esser et al., 2021) to transform the continuous video stream into discrete visual tokens. Specifically, given a continuous video sequence $\mathbf{V}_{t-H:t+T}$, we first partition the sequence into $N$ non-overlapping chunks, where $N = (H+T)/t_{\text{chunk}} + 1$ and $t_{\text{chunk}}$ denotes the chunk duration. For each chunk, we employ a VQ tokenizer to discretize the first visual observation into discrete visual tokens, yielding:

$$V_q = \{V_q^0, V_q^1, \ldots, V_q^{N-1}\}, \tag{2}$$

where $V_q^j$ represents the quantized visual tokens corresponding to the initial frame of the $j$-th chunk.

**Action Modeling.** We adopt FAST (Pertsch et al., 2025) to discretize continuous action sequences into discrete action tokens. Given the waypoint sequence $\mathbf{W}_{t-H:t+T}$, we convert absolute coordinates to ego-relative coordinates to obtain the action sequence $V_{t-H:t+T}$, which is split into chunks and encodes $V_{t-H:t+T}$ into discrete tokens using the FAST action tokenizer, yielding:

$$A_q = \{A_q^0, A_q^1, \ldots, A_q^{N-1}\}, \tag{3}$$

where each $A_q^j$ corresponds to the action tokens within the $j$-th temporal chunk.

**Language Modeling.** We utilize the Emu3 tokenizer (Wang et al., 2024b) to discretize the driving command $\mathbf{CMD}_{t-H:t+T}$ into language tokens:

$$L_q = \{L_q^0, L_q^1, \ldots, L_q^{N-1}\}, \tag{4}$$

where each $L_q^j$ corresponds to the $j$-th language token.

**Future Latent Prediction.** We strategically insert a learnable query token $Q$ into the input token sequence, and the hidden state at the query position is fed to an MLP to predict the future latent representation $\hat{\mathbf{z}}_2$ during the forward pass.

**Training Objective.** Since all modality signals are transformed into discrete tokens drawn from a shared vocabulary, we construct the world model's input sequence as:

$$\mathcal{S} = [L_q^0, V_q^0, Q, A_q^0, L_q^1, V_q^1, A_q^1, L_q^2, \ldots, V_q^{N-1}, A_q^{N-1}], \tag{5}$$

To prevent information leakage, causal masking is applied such that each token can only attend to preceding tokens in the sequence. The final training objective consists of three terms:

$$\mathcal{L}_{\text{world}} = \sum_{j=0}^{N-1} \text{CE}(\hat{L}_q^j, L_q^j) + \sum_{j=0}^{N-1} \text{CE}(\hat{A}_q^j, A_q^j) + \lambda_1 \sum_{j=0}^{N-1} \text{CE}(\hat{v}_q^j, v_q^j) + \lambda_2 \|\hat{\mathbf{z}}_2 - \mathbf{z}_2\|_2^2. \tag{6}$$

Here, the first term is the language tokens prediction loss, the second term is the action prediction loss, the third term is the visual tokens prediction loss, and the last term enforces consistency between the predicted latent representation and the ground-truth compressed feature extracted from the future video sequence. $\lambda_1$ and $\lambda_2$ are the vision token and latent representation loss weights.

### 3.4 Co-Fine-Tuning

In the co-fine-tuning phase, the world model and the action expert are jointly optimized to achieve continuous and reliable action prediction.

We integrate the backbone and the action expert via a joint attention mechanism (Black et al., 2024). The historical actions $V_{t-H:t}$, the driving commands $\mathbf{CMD}_{t-H:t}$ and the predicted latent representation $\hat{\mathbf{z}}_2$ from the backbone are concatenated as the sequence input to the attention module. After interacting with the backbone context, the hidden state at the action token positions is decoded via an MLP head.

To verify cross-paradigm compatibility, we instantiate our method on the two dominant action model families: selection-based (Li et al., 2024) and generative-model-based (Lipman et al., 2022).

For the generative-model training objectives, including flow matching, we have the following:

$$\mathcal{L}_{\text{cotrain}} = |\hat{v} - v|_2^2 + \lambda_3 |\hat{z} - z|_2^2 \tag{7}$$

The first term is the flow-matching vector field regression loss, and the second term is the latent prediction loss, and $\lambda_3$ is the latent loss weight.

The selection-based training objectives are as follows:

$$\mathcal{L}_{\text{cotrain}} = \text{KL}\left(p_{\text{pred}} \parallel p_{\text{gt}}\right) + \lambda_3 \|\hat{z} - z\|_2^2, \tag{8}$$

where $p_{\text{pred}}$ is the predicted score for each trajectory in an offline trajectory vocabulary, $p_{\text{gt}}$ is the ground-truth score computed by comprehensively considering the expert trajectory as well as the metrics of the evaluation and a KL divergence loss is employed to supervise the predicted score distribution.

## 4 Experiments

### 4.1 Datasets

We evaluate our approach on the NAVSIM v1 (Dauner et al., 2024) and v2 (Cao et al., 2025) benchmarks, providing closed-loop evaluation protocols and multi-dimensional quantitative metrics.

**NAVSIM v1.** NAVSIM v1 measures driving capability from five core dimensions: No at-fault Collision (NC), Drivable Area Compliance (DAC), Time-To-Collision (TTC), Comfort (C.), and Ego Progress (EP). The overall performance is aggregated into the *Predictive Driver Model Score (PDMS)* via a weighted multiplicative formulation:

$$\text{PDMS} = \text{NC} \times \text{DAC} \times \frac{5 \times \text{EP} + 5 \times \text{TTC} + 2 \times \text{C.}}{12}. \tag{9}$$

**NAVSIM v2.** As an extended and refined version, NAVSIM v2 (Cao et al., 2025) introduces more fine-grained evaluation dimensions to better align with real-world driving requirements. In addition to the core metrics inherited from v1, it adds Driving Direction Compliance (DDC), Traffic Light Compliance (TLC), Lane Keeping (LK), History Comfort (HC), and Extended Comfort (EC). The corresponding comprehensive metric is the *Extended Predictive Driver Model Score (EPDMS)*, defined as:

$$\text{EPDMS} = \text{NC} \times \text{DAC} \times \text{DDC} \times \text{TLC} \times \frac{5 \times \text{EP} + 5 \times \text{TTC} + 2 \times \text{LK} + 2 \times \text{HC} + 2 \times \text{EC}}{16}. \quad (10)$$

### 4.2 Implementation Details

Our model backbone is based on Emu3 (Wang et al., 2024b), a unified multimodal autoregressive model.

**VideoVAEPlus Fine-tuning.** We initialize from the 4-channel VideoVAEPlus (Xing et al., 2024) pre-trained model and fine-tune it on the NuPlan dataset (Caesar et al., 2021). For each video segment, we uniformly sample 8 frames resized to $216 \times 216$ and use Qwen3.6-plus to generate a short scene description. The training is performed on 4 PPUs (Alibaba Cloud, 2026) with a per-PPU batch size of 1. The weights of the adversarial loss and the KL loss are set to 0.5 and 1e-6 respectively, and the adversarial loss is enabled after 50K steps. We train the model for 100K steps, enabling it to learn compact representations from the videos.

**Pre-training.** During the pre-training stage, we utilize the OpenScenes dataset (Contributors, 2023) comprising over 120 hours of driving video recordings which are segmented into clips of 20.0 s. During training, for each clip, we randomly sample a frame sequence with a future horizon $T = 4.0$ s and a history horizon $H = 1.0$ s. Given the chunk duration $t_{\text{chunk}} = 1.0$ s, the number of non-overlapping chunks num $N$ is 6. The training configuration adopts a batch size of 4 across two nodes, each equipped with 16 PPUs (Alibaba Cloud, 2026). We set the learning rate to $2.2 \times 10^{-4}$ with a maximum of 4000 training steps. The loss weighting coefficients are configured as $\lambda_1 = 0.5$ and $\lambda_2 = 0.1$.

**Co-fine-tuning.** During the co-fine-tuning phase, we fine-tune the pre-trained model on the NAVSIM trainval dataset (Dauner et al., 2024) which comprises over 100,000 frames with ground-truth ego-vehicle trajectory annotations. Consistent with the pre-training configuration, the future horizon $T$ and historical horizon $H$ are set to 4 s and 1 s, respectively. We train for 4000 steps with a total batch size of 96. The learning rate is set to $5 \times 10^{-5}$, with a latent prediction loss weight $\lambda_3 = 0.1$.

### 4.3 Comparison with State-of-the-Art Methods

We benchmark HyWorldVLA against representative pixel-based world model (WM) methods, latent-based world model methods, as well as several canonical end-to-end and VLA-based baselines. As shown in Table 1 and Table 2, our approach achieves state-of-the-art (SOTA) performance on both v1 and v2 benchmarks, significantly outperforming the competitive pixel-based WM baselines such as DriveVLA-W0 (Li et al., 2025c) and the latent-based WM baselines including Latent-WAM (Wang et al., 2026). This performance gain can be ascribed to our hybrid world modeling paradigm, which reconciles the fine-grained spatiotemporal modeling capability of pixel-based world models with the intrinsic high-level semantic representation capacity of latent world models.

### 4.4 Ablation Analysis

**Ablation on Components.** Table 3 presents the ablation results for each component of HyWorldVLA. Removing the pixel-regression ability (Pure LWM) yields the lowest score (87.5), confirming the crucial role of its spatiotemporal modeling capability in understanding the environment and generating accurate actions. Removing the latent representational capacity (Pure WAM) also leads to a significant drop (89.9), indicating its positive effect on learning high-level semantics. Ablating the language guidance from the latent module decreases the score to 90.3, demonstrating that language effectively guides the model toward planning-relevant semantics. Removing the latent conditioning in the action expert model drops the score to 90.3, suggesting that explicitly leveraging predicted future information benefits more accurate action generation. Eliminating the latent supervision during

Table 1: Comparison with state-of-the-art method on the NAVSIM v1. Abbreviations: 1x Cam (single front-view camera), Nx Cam (surround-view cameras), L (LiDAR).

| Method | Ref | Sensors | NC↑ | DAC↑ | TTC↑ | C.↑ | EP↑ | PDMS↑ |
|---|---|---|---|---|---|---|---|---|
| Human | - | - | 100 | 100 | 100 | 99.9 | 87.5 | 94.8 |
| *End-to-end-based Methods* | | | | | | | | |
| UniAD (Hu et al., 2023b) | CVPR'23 | 6x Cam | 97.8 | 91.9 | 92.9 | **100.0** | 78.8 | 83.4 |
| TransFuser (Prakash et al., 2021) | TPAMI'23 | 3x Cam + L | 97.7 | 92.8 | 92.8 | **100.0** | 79.2 | 84.0 |
| PARA-Drive (Weng et al., 2024) | CVPR'24 | 6x Cam | 97.9 | 92.4 | 93.0 | 99.8 | 79.3 | 84.0 |
| Hydra-MDP (Li et al., 2024) | arXiv'24 | 3x Cam + L | 98.3 | 96.0 | 94.6 | **100.0** | 78.7 | 86.5 |
| DiffusionDrive (Liao et al., 2025) | CVPR'25 | 3x Cam + L | 98.2 | 96.2 | 94.7 | **100.0** | 82.2 | 88.1 |
| *VLA-based Methods* | | | | | | | | |
| AutoVLA (Zhou et al., 2025b) | NeurIPS'25 | 3x Cam | 98.4 | 95.6 | **98.0** | 99.9 | 81.9 | 89.1 |
| *Pixel-based Word Model Methods* | | | | | | | | |
| WoTE (Li et al., 2025d) | ICCV'25 | 3x Cam + L | 98.5 | 96.8 | 94.4 | 99.9 | 81.9 | 88.3 |
| FSDrive (Zeng et al., 2026) | NeurIPS'25 | 1x Cam | 98.2 | 93.8 | 93.3 | 99.9 | 80.1 | 85.1 |
| ReSim (Yang et al., 2026a) | NeurIPS'25 | 1x Cam | - | - | - | - | - | 86.6 |
| PWM (Zhao et al., 2026) | NeurIPS'25 | 1x Cam | 98.6 | 95.9 | 95.4 | **100.0** | 81.8 | 88.1 |
| CoWorld-VLA (Huang et al., 2026) | arXiv'26 | 1x Cam | **99.2** | 96.8 | 96.6 | **100.0** | 83.6 | 89.8 |
| DriveLaW (Xia et al., 2026) | CVPR'26 | 1x Cam | 99.0 | 97.1 | 96.7 | **100.0** | 81.3 | 89.1 |
| DriveVLA-W0 (Li et al., 2025c) | ICLR'26 | 1x Cam | 98.7 | **99.1** | 95.3 | 99.3 | 83.3 | 90.2 |
| *Latent-based World Model Methods* | | | | | | | | |
| LAW (Li et al., 2025b) | ICLR'25 | 1x Cam | 96.4 | 95.4 | 88.7 | 99.9 | 81.7 | 84.6 |
| World4Drive (Zheng et al., 2025) | ICCV'25 | 1x Cam | 97.4 | 94.3 | 92.8 | **100.0** | 79.9 | 85.1 |
| Epona (Zhang et al., 2025) | ICCV'25 | 1x Cam | 97.9 | 95.1 | 93.8 | 99.9 | 80.4 | 86.2 |
| DreamerAD (Yang et al., 2026c) | arXiv'26 | 1x Cam | 98.0 | 97.2 | 94.3 | **100.0** | 83.1 | 88.7 |
| OneVL (Lu et al., 2026b) | arXiv'26 | 1x Cam | - | - | - | - | - | 88.8 |
| WorldRFT (Yang et al., 2026b) | AAAI'26 | 1x Cam | 97.8 | 96.8 | 94.0 | **100.0** | 81.7 | 87.8 |
| ResWorld (Zhang et al., 2026) | ICLR'26 | 4x Cam + L | 98.9 | 96.5 | 95.6 | **100.0** | 83.1 | 89.0 |
| Ours | - | 1x Cam | 98.6 | 98.9 | 95.1 | 99.3 | **84.6** | **90.6** |

Table 2: Comparison with state-of-the-art methods on the NAVSIM v2 with extended metrics.

| Method | NC↑ | DAC↑ | DDC↑ | TLC↑ | EP↑ | TTC↑ | LK↑ | HC↑ | EC↑ | EPDMS↑ |
|---|---|---|---|---|---|---|---|---|---|---|
| *End-to-end-based Methods* | | | | | | | | | | |
| TransFuser (Prakash et al., 2021) | 96.9 | 89.9 | 97.8 | 99.7 | 87.1 | 95.4 | 92.4 | 98.3 | 87.2 | 76.7 |
| HydraMDP++ (Li et al., 2025a) | 97.2 | 97.5 | 99.4 | 99.6 | 83.1 | 96.5 | 94.4 | 98.2 | 70.9 | 81.4 |
| DriveSuprim (Yao et al., 2026) | 97.5 | 96.5 | 99.4 | 99.6 | 88.4 | 96.6 | 95.5 | 98.3 | 77.0 | 83.1 |
| ARTEMIS (Feng et al., 2025) | 98.3 | 95.1 | 98.6 | 99.8 | 81.5 | 97.4 | 96.5 | 98.3 | - | 83.1 |
| DiffusionDriveV2 (Zou et al., 2025) | 97.7 | 96.6 | 99.2 | 99.8 | **88.9** | 97.2 | 96.0 | 97.8 | **91.0** | 85.5 |
| *VLA-based Methods* | | | | | | | | | | |
| ReCogDrive (Li et al., 2025e) | 98.3 | 95.2 | 98.3 | 99.8 | 87.1 | 97.5 | 96.6 | **99.5** | 86.5 | 83.6 |
| *Pixel-based Word Model Methods* | | | | | | | | | | |
| DriveVLA-W0 (Li et al., 2025c) | 98.5 | **99.1** | 98.0 | 99.7 | 86.4 | 98.1 | 93.2 | 97.9 | 58.9 | 86.1 |
| ExploreVLA (Sheng et al., 2026) | **98.8** | 96.2 | **99.6** | 99.8 | 87.1 | **98.2** | **97.8** | 98.3 | 86.8 | 88.8 |
| *Latent-based World Model Methods* | | | | | | | | | | |
| DriveWorld-VLA (Liu et al., 2026) | 98.6 | **99.1** | **99.6** | 99.8 | 87.4 | 97.9 | 97.0 | 97.9 | 78.6 | 86.8 |
| DreamerAD (Yang et al., 2026c) | 98.6 | **99.1** | **99.6** | 99.8 | 87.4 | 97.9 | 97.0 | 97.9 | 78.6 | 86.8 |
| Latent-WAM (Wang et al., 2026) | 98.0 | 97.2 | 99.5 | 99.8 | 87.8 | 97.4 | 97.5 | 98.3 | 72.4 | 87.7 |
| Ours | **98.8** | 98.2 | 99.5 | **99.9** | 87.8 | 98.1 | 96.2 | 98.3 | 77.4 | **89.7** |

co-fine-tuning results in a score of 90.2, highlighting the importance of preserving pre-trained latent semantics and preventing over-optimization.

**Ablation on Loss Weights in Pre-training.** As shown in Table 4, We first fix $\lambda_2 = 0.1$ and examine the effect of $\lambda_1$. When $\lambda_1$ increases from 0.1 to 0.5 and 1.0, the score improves from 90.5 to 90.6

Table 3: Ablation study on the NAVSIM benchmark.

| Config | NC↑ | DAC↑ | TTC↑ | C.↑ | EP↑ | PDMS↑ |
|---|---|---|---|---|---|---|
| Pure WAM | 98.3 | 98.6 | 94.5 | 99.6 | 84.0 | 89.9 |
| Pure LWM | 97.2 | 97.8 | 92.0 | 98.9 | 82.4 | 87.5 |
| w/o language guidance in latent | 98.5 | 98.8 | 94.7 | 99.6 | 84.5 | 90.3 |
| w/o latent condition in action expert | **98.6** | 98.7 | 94.9 | 99.3 | 84.3 | 90.3 |
| w/o latent supervision in co-fine-tuning | **98.6** | **98.9** | **95.3** | **99.8** | 83.3 | 90.2 |
| Ours | **98.6** | **98.9** | 95.1 | 99.3 | **84.6** | **90.6** |

and drops significantly to 89.8. This indicates that the weight for visual tokens prediction should not be too high. When we fix $\lambda_1 = 0.5$ and increase $\lambda_2$ from 0.1 to 0.2, 0.5 and 1.0, the score decreases from 90.6 to 90.5 and then continues to drop to 90.3 and 90.2, respectively. This demonstrates that simultaneously introducing visual tokens prediction ($\lambda_1$ = 0.5) and latent motion ($\lambda_2$ = 0.1) during the pre-training phase most effectively guides the trajectory.

**Ablation on Loss Weights in co-fine-tuning.** As shown in Table 5, starting from $\lambda_3 = 0.05$, we observe that increasing it to 0.1 raises the score from 90.5 to 90.6, indicating that appropriate supervision prevents performance degradation caused by semantic collapse of the latent space. However, when $\lambda_3$ is further increased to 0.2 and 1.0, the score drops to 90.0 and 89.8, respectively, suggesting that overly strong supervision hinders the model from autonomously learning the latent semantics most relevant to action generation.

Table 4: Ablation study of the loss weights in pre-training stage.

| $\lambda_1$ | $\lambda_2$ | NC↑ | DAC↑ | TTC↑ | C.↑ | EP↑ | PDMS↑ |
|---|---|---|---|---|---|---|---|
| 0.1 | 0.1 | 98.6 | 98.7 | 94.9 | 99.6 | 84.6 | 90.5 |
| 0.5 | 0.1 | 98.6 | 98.9 | 95.1 | 99.3 | 84.6 | **90.6** |
| 1.0 | 0.1 | 98.5 | 98.5 | 94.9 | 99.6 | 83.5 | 89.8 |
| 0.5 | 0.2 | 98.6 | 98.9 | 95.0 | 99.5 | 84.4 | 90.5 |
| 0.5 | 0.5 | 98.6 | 98.6 | 94.5 | 99.4 | 84.9 | 90.3 |
| 0.5 | 1.0 | 98.6 | 98.6 | 94.7 | 99.5 | 84.4 | 90.2 |

Table 5: Ablation study of the loss weights in co-fine-tuning stage.

| $\lambda_3$ | NC↑ | DAC↑ | TTC↑ | C.↑ | EP↑ | PDMS↑ |
|---|---|---|---|---|---|---|
| 0.05 | 98.5 | 98.8 | 95.0 | 99.3 | 84.6 | 90.5 |
| 0.1 | 98.6 | 98.9 | 95.1 | 99.3 | 84.6 | **90.6** |
| 0.2 | 98.6 | 98.7 | 94.6 | 99.4 | 83.9 | 90.0 |
| 1.0 | 98.6 | 98.3 | 95.0 | 99.4 | 83.4 | 89.8 |

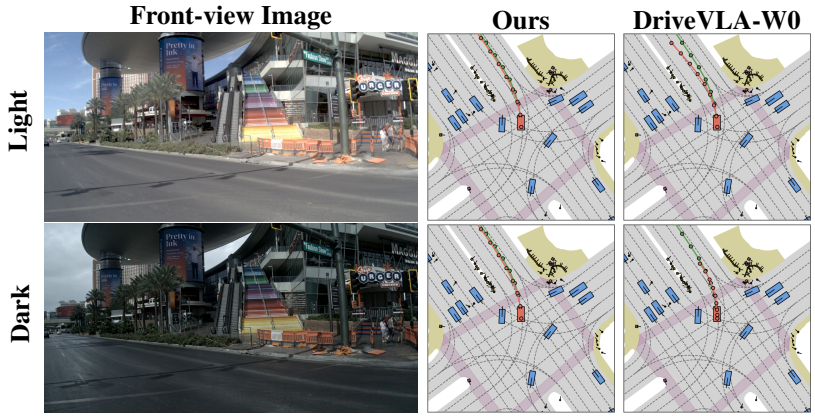
### 4.5 Scene Noise Robustness Analysis

**Quantitative Analysis.** To further analyze the behavior of HyWorldVLA under the noisy conditions, a noise robustness test set was built by collecting 655 cases from the Openscene test dataset (Contributors, 2023), all of which contain non-uniform noise due to rain and fog. Several methods and configurations are compared in terms of performance: WoTE (Li et al., 2025d), DriveLaW (Xia et al., 2026), DriveVLA-W0 (Li et al., 2025c), HyWorldVLA(Pure WM), HyWorldVLA(w/o latent supervision) and HyWorldVLA. As shown in Table 6, HyWorldVLA achieves the highest performance of 86.9 on corrupted cases, while WoTE (Li et al., 2025d), DriveLaW (Xia et al., 2026) and DriveVLA-W0 (Li et al., 2025c) obtain only 60.7, 67.5 and 61.2, respectively, far below that of HyWorldVLA, demonstrating the superior robustness of our method. Pure WAM and w/o latent supervision in co-fine-tuning configurations obtain 70.0 and 73.2, which further clarifies that latent representation design of the HyWorldVLA yields this robust performance in noisy scenarios.

**Qualitative Analysis.** For a qualitative robustness assessment,we systematically modulate front-view illumination, transitioning from sunny to overcast conditions. As illustrated in Fig. 3a, HyWorldVLA maintains consistent behavioral patterns largely invariant to lighting variations, whereas DriveVLA-W0 exhibits significant instability. Furthermore, Fig. 3b showcases two additional cases drawn from our proposed noise robustness test set, featuring non-uniform visual degradations. Under these challenging conditions, HyWorldVLA continues to execute correct driving maneuvers, while DriveVLA-W0 fails to respond appropriately. Additional cases are in Appendix C (Figs. 11-13).

Table 6: Scene noise robustness comparison with other methods on the NAVSIM benchmark.

| Method/Config | NC↑ | DAC↑ | TTC↑ | C.↑ | EP↑ | PDMS↑ |
|---|---|---|---|---|---|---|
| WoTE (Li et al., 2025d) | 98.9 | 76.0 | 64.4 | **100.0** | 64.4 | 60.7 |
| DriveLaW (Xia et al., 2026) | 99.4 | 75.7 | 82.9 | **100.0** | 69.5 | 67.5 |
| DriveVLA-W0 (Li et al., 2025c) | 93.4 | 76.2 | 83.1 | 98.6 | 59.8 | 61.2 |
| Pure WAM | 99.1 | 75.7 | 94.3 | 99.8 | 67.8 | 70.0 |
| w/o latent supervision in co-fine-tuning | 99.2 | 78.5 | 94.8 | 99.7 | 70.9 | 73.2 |
| Ours | **99.7** | **91.1** | **96.9** | 99.2 | **84.4** | **86.9** |




(a) Turn right under the fluctuating illumination.




(b) Turn left and go straight under the non-uniform noise from rain and fog.

NPC Ego Car Lane Center Line GT Traj. Pred Traj.

Figure 3: **Hybrid World modeling improves the robustness to the scene noise.** (a) DriveVLA-W0 suffers from severe behavioral variance under illumination changes, while our approach retains remarkable consistency. (b) DriveVLA-W0 exhibits overly conservative behavior that sacrifices traffic efficiency, while our method retains accurate driving with human-aligned trajectories.

## 5 Conclusion

We present a hybrid autonomous driving world model that combines the fine-grained spatiotemporal modeling of pixel-based methods with the noise robustness of latent variants. Pixel-level reconstruction preserves detailed spatial and dynamic information, while latent prediction in the VAE space inherently improves noise resilience. Experiments on NAVSIM v1/v2 show our method achieves state-of-the-art performance, with further validation on rain/fog-corrupted inputs. Limitations and future work include extending the monocular-based framework to multi-modal sensors and strengthening latent dynamics coupling with critical traffic elements.

## AI USE STATEMENT

In this work, we used generative AI tools for creating or editing software code. We have not used generative AI tools for generating synthetic data sets, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, cleaning and reformatting dataset, supporting qualitative and thematic data analysis and interpreting results. Developing theoretical models or conceptual frameworks, formulating mathematical claims, proving mathematical claims, writing of proofs are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## REFERENCES


Alibaba Cloud. Ppu introduction. [Online], 2026. https://help.aliyun.com/zh/document_detail/2864586.html.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. $\pi_0$: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

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

Wei Cao, Marcel Hallgarten, Tianyu Li, Daniel Dauner, Xunjiang Gu, Caojun Wang, Yakov Miron, Marco Aiello, Hongyang Li, Igor Gilitschenski, et al. Pseudo-simulation for autonomous driving. *arXiv preprint arXiv:2506.04218*, 2025.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

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

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, et al. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *arXiv preprint arXiv:2406.15349*, 2024.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Renju Feng, Ning Xi, Duanfeng Chu, Rukang Wang, Zejian Deng, Anzheng Wang, Liping Lu, Jinxiang Wang, and Yanjun Huang. Artemis: Autoregressive end-to-end trajectory planning with mixture of experts for autonomous driving. *IEEE Robotics and Automation Letters*, 11(1):226–233, 2025.

Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Dingkang Liang, Chong Zhang, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. *arXiv preprint arXiv:2503.19755*, 2025.

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.

Tianshuai Hu, Xiaolu Liu, Song Wang, Yiyao Zhu, Ao Liang, Lingdong Kong, Guoyang Zhao, Zeying Gong, Jun Cen, Zhiyu Huang, et al. Vision-language-action models for autonomous driving: Past, present, and future. *arXiv preprint arXiv:2512.16760*, 2025.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17853–17862, 2023b.

Minqing Huang, Yujiao Xiang, Zihan Liang, Jiajie Huang, Jingqi Wang, Zhi Xu, Feiyang Tan, Hangning Zhou, Mu Yang, and Gong Che. Coworld-vla: Thinking in a multi-expert world model for autonomous driving. *arXiv preprint arXiv:2605.10426*, 2026.

Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, et al. EMMA: End-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2410.23262*, 2024.

Fan Jia, Weixin Mao, Yingfei Liu, Yucheng Zhao, Yuqing Wen, Chi Zhang, Xiangyu Zhang, and Tiancai Wang. Adriver-i: A general world model for autonomous driving. *arXiv preprint arXiv:2311.13549*, 2023.

Bo Jiang, Shaoyu Chen, Qian Zhang, Wenyu Liu, and Xinggang Wang. AlphaDrive: Unleashing the power of VLMs in autonomous driving via reinforcement learning and reasoning. *arXiv preprint arXiv:2503.07608*, 2025.

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

Yingyan Li, Lue Fan, Jiawei He, Yuqi Wang, Yuntao Chen, Zhaoxiang Zhang, and Tieniu Tan. Enhancing end-to-end autonomous driving with latent world model. In *International Conference on Learning Representations*, volume 2025, pp. 42942–42959, 2025b.

Yingyan Li, Shuyao Shang, Weisong Liu, Bing Zhan, Haochen Wang, Yuqi Wang, Yuntao Chen, Xiaoman Wang, Yasong An, Chufeng Tang, Lu Hou, Lue Fan, and Zhaoxiang Zhang. Drivevla-w0: World models amplify data scaling law in autonomous driving. *arXiv preprint arXiv:2510.12796*, 2025c.

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

Yongkang Li, Kaixin Xiong, Xiangyu Guo, Fang Li, Sixu Yan, Gangwei Xu, Lijun Zhou, Long Chen, Haiyang Sun, Bing Wang, Kun Ma, Guang Chen, Hangjun Ye, Wenyu Liu, and Xinggang Wang. Recogdrive: A reinforced cognitive framework for end-to-end autonomous driving. *arXiv preprint arXiv:2506.08052*, 2025e.

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

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

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

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

Jinghui Lu, Jiayi Guan, Zhijian Huang, Jinlong Li, Guang Li, Lingdong Kong, Yingyan Li, Han Wang, Shaoqing Xu, Yuechen Luo, Fang Li, Chenxu Dang, Junli Wang, Tao Xu, Jing Wu, Jianhua Wu, Xiaoshuai Hao, Wen Zhang, Tianyi Jiang, Lingfeng Zhang, Lei Zhou, Yingbo Tang, Jie Wang, Yinfeng Gao, Feiyang Jia, Lin Liu, Yigu Ge, Hanbing Li, Yuannan Shen, Jianwei Cui, Hongwei Xie, and Bing Wang. Xiaomi onevl: One-step latent reasoning and planning with vision-language explanation. *arXiv preprint arXiv:2604.18486*, 2026a.

Jinghui Lu, Jiayi Guan, Zhijian Huang, Jinlong Li, Guang Li, Lingdong Kong, Yingyan Li, Han Wang, Shaoqing Xu, Yuechen Luo, et al. Xiaomi onevl: One-step latent reasoning and planning with vision-language explanation. *arXiv preprint arXiv:2604.18486*, 2026b.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 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*, pp. 7077–7087, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Zihao Sheng, Xin Ye, Jingru Luo, Sikai Chen, and Liu Ren. Explorevla: Dense world modeling and exploration for end-to-end autonomous driving. *arXiv preprint arXiv:2604.02714*, 2026.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. DriveLM: Driving with graph visual question answering. In *European conference on computer vision*, pp. 256–274. Springer, 2024.

Shuhan Tan, Kashyap Chitta, Yuxiao Chen, Ran Tian, Yurong You, Yan Wang, Wenjie Luo, Yulong Cao, Philipp Krahenbuhl, Marco Pavone, and Boris Ivanovic. Latent chain-of-thought world modeling for end-to-end driving. *arXiv preprint arXiv:2512.10226*, 2025. Accepted to CVPR 2026.

Sifan Tu, Xin Zhou, Dingkang Liang, Xingyu Jiang, Yumeng Zhang, Xiaofan Li, and Xiang Bai. The role of world models in shaping autonomous driving: A comprehensive survey. *arXiv preprint arXiv:2502.10498*, 2025.

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

Wenhai Wang, Jiangwei Xie, ChuanYang Hu, Haoming Zou, Jianan Fan, Wenwen Tong, Yang Wen, Silei Wu, Hanming Deng, Zhiqi Li, et al. DriveMLM: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *arXiv preprint arXiv:2312.09245*, 2023.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European conference on computer vision*, pp. 55–72. Springer, 2024a.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024b.

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

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

Yazhou Xing, Yang Fei, Yingqing He, Jingye Chen, Jiaxin Xie, Xiaowei Chi, and Qifeng Chen. Large motion video autoencoding with cross-modal video vae. *arXiv preprint arXiv:2412.17805*, 2024.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. DriveGPT4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 9(10):8186–8193, 2024.

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

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

Pengxuan Yang, Yupeng Zheng, Deheng Qian, Zebin Xing, Qichao Zhang, Linbo Wang, Yichen Zhang, Shaoyu Guo, Zhongpu Xia, Qiang Chen, et al. Dreamerad: Efficient reinforcement learning via latent world model for autonomous driving. *arXiv preprint arXiv:2603.24587*, 2026c.

Wenhao Yao, Zhenxin Li, Shiyi Lan, Zi Wang, Xinglong Sun, Jose M Alvarez, and Zuxuan Wu. Drivesuprim: Towards precise trajectory selection for end-to-end planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11910–11918, 2026.

Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, Xing Wei, and Ning Guo. Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. *arXiv preprint arXiv:2505.17685*, 2025. Accepted to NeurIPS 2025 as Spotlight.

Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, and Xing Wei. Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. *Advances in Neural Information Processing Systems*, 38:67299–67318, 2026.

Jinqing Zhang, Zehua Fu, Zelin Xu, Wenying Dai, Qingjie Liu, and Yunhong Wang. Resworld: Temporal residual world model for end-to-end autonomous driving. *arXiv preprint arXiv:2602.10884*, 2026.

Kaiwen Zhang, Zhenyu Tang, Xiaotao Hu, Xingang Pan, Xiaoyang Guo, Yuan Liu, Jingwei Huang, Li Yuan, Qian Zhang, Xiao-Xiao Long, 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.

Zhida Zhao, Talas Fu, Yifan Wang, Lijun Wang, and Huchuan Lu. From forecasting to planning: Policy world model for collaborative state-action prediction. *Advances in Neural Information Processing Systems*, 38:134585–134611, 2026.

Wenzhao Zheng, Zetian Xia, Yuanhui Huang, Sicheng Zuo, Jie Zhou, and Jiwen Lu. Doe-1: Closed-loop autonomous driving with large world model. *arXiv preprint arXiv:2412.09627*, 2024.

Yupeng Zheng, Pengxuan Yang, Zebin Xing, Qichao Zhang, Yuhang Zheng, Yinfeng Gao, Pengfei Li, Teng Zhang, Zhongpu Xia, Peng Jia, et al. World4drive: End-to-end autonomous driving via intention-aware physical latent world model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 28632–28642, 2025.

Xingcheng Zhou, Xuyuan Han, Feng Yang, Yunpu Ma, Volker Tresp, and Alois Knoll. Opendrivevla: Towards end-to-end autonomous driving with large vision language action model. *arXiv preprint arXiv:2503.23463*, 2025a.

Zewei Zhou, Tianhui Cai, Seth Z. Zhao, Yun Zhang, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. *arXiv preprint arXiv:2506.13757*, 2025b. Accepted by NeurIPS 2025.

Jialv Zou, Shaoyu Chen, Bencheng Liao, Zhiyu Zheng, Yuehao Song, Lefei Zhang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Diffusiondrivev2: Reinforcement learning-constrained truncated diffusion modeling in end-to-end autonomous driving. *arXiv preprint arXiv:2512.07745*, 2025.

# APPENDIX

**A Overall** **15**

**B More Experiments** **15**

B.1 Flow Matching Based Experiments . . . . . 15

B.2 Ablation Study For The Scene Noise Robustness . . . . . 15

**C Visualization** **16**

C.1 Future Image Generation . . . . . 16

C.2 Planning Trajectory . . . . . 17

C.3 Failure Case . . . . . 18

C.4 More Scene Noise Robustness Case . . . . . 18

## A OVERALL

This appendix contains supplementary materials that accompany the main paper. § B reports additional experiments, which include the comparison of the flow matching action expert on the NAVSIM benchmark, as well as an ablation study of the latent loss weights for the scene noise robustness. Further qualitative findings are presented in § C, including images produced by our world model, visualizations of challenging cases and qualitative analysis of illumination interference and blur noise cases.

## B MORE EXPERIMENTS

### B.1 FLOW MATCHING BASED EXPERIMENTS

In order to verify the versatility of our approach on different types of action models, in addition to the anchor selective action model method, we also conduct comparative experiments with the flow-matching action model as shown in Table 7.

Table 7: Flow matching as action expert version comparison on the NAVSIM benchmark.

| Method | NC↑ | DAC↑ | TTC↑ | C.↑ | EP↑ | PDMS↑ |
|---|---|---|---|---|---|---|
| DriveVLA-W0 (fm) | 98.4 | 95.3 | 95.2 | **100.0** | 80.9 | 87.2 |
| Ours (fm) | **98.8** | **96.3** | **95.5** | **100.0** | **82.9** | **88.76** |

### B.2 ABLATION STUDY FOR THE SCENE NOISE ROBUSTNESS

To further clarify how latent representation contributes to noise robustness, we perform an ablation study on $\lambda_3$. Table 8 shows that larger $\lambda_3$ consistently boosts model performance on noisy data, validating the beneficial role of latent representation in noise robustness. Nevertheless, unbounded increase of $\lambda_3$ degrades performance on normal cases, making the trade-off between these two capabilities a critical future research direction.

Table 8: Ablation study of the latent loss weights on the noise robustness test.

| $\lambda_3$ | NC↑ | DAC↑ | TTC↑ | C.↑ | EP↑ | PDMS↑ |
|---|---|---|---|---|---|---|
| 0.05 | 99.5 | 88.7 | 96.6 | 99.1 | 82.4 | 84.52 |
| 0.1 | 99.7 | 91.1 | 96.9 | 99.2 | 84.4 | 86.87 ↑+2.35 |
| 0.2 | 99.1 | 94.8 | 94.5 | 99.4 | 88.1 | 89.45 ↑+2.58 |
| 1.0 | 98.8 | 98.3 | 95.1 | 100.0 | 90.8 | 92.94 ↑+3.49 |

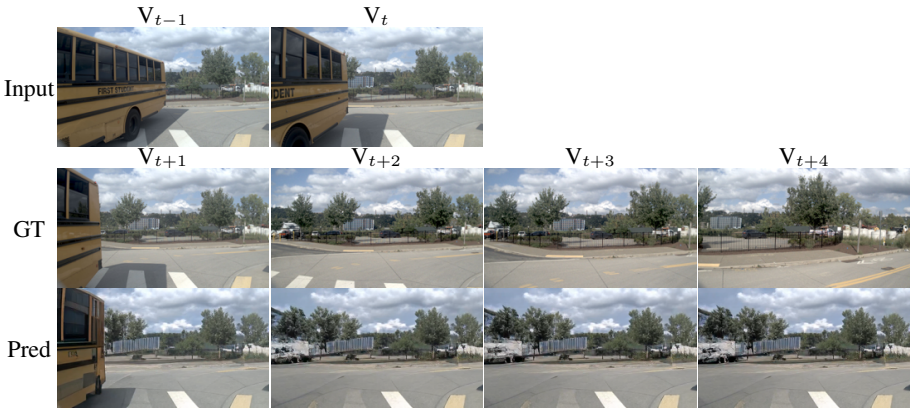
# C VISUALIZATION

## C.1 FUTURE IMAGE GENERATION

Our model inherits the strong future frame prediction capability of the World Action Model. As shown in Figure 4a, with only a single front-view camera, our model generates the central double yellow line absent from the input frames. Moreover, our model accurately forecasts future maneuvers of the ego vehicle and obstacles in the highly interactive scenario, as shown in Figure 4b.




(a) Right turn in the intersection with a limited front camera view.




(b) Meeting an oncoming bus on a narrow road.

Figure 4: **Future image generation.** (a) During right turns, despite the absence of lane markings in front-view camera inputs, our model reliably forecasts future lane markings, evidencing its learned world knowledge. (b) Our model correctly forecasts the motion pattern of the oncoming bus and the ego car's corresponding rightward yielding action, evidencing its complex scene understanding capability.

### C.2 PLANNING TRAJECTORY

In this section, we conduct a qualitative case analysis to gain deeper insights into our model's behavior and validate our key findings. Specifically, we compare our Hybrid World Model against the baseline DriveVLA-W0 across multiple challenging scenarios to demonstrate, as shown in Figure 5, Figure 6 and Figure 7.

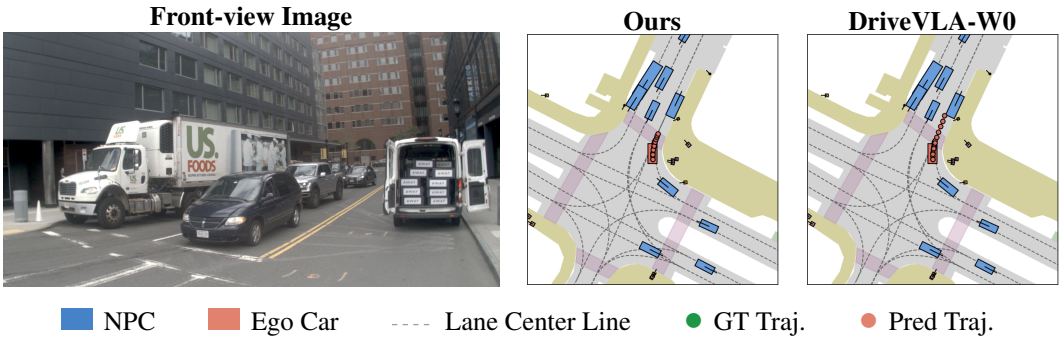



Figure 5: **Turn right and stop behind a truck with its trunk open.** DriveVLA-W0 baseline fails in trying to nudge a stopping truck with its trunk open. Our method stops and keep a fit distance to it.

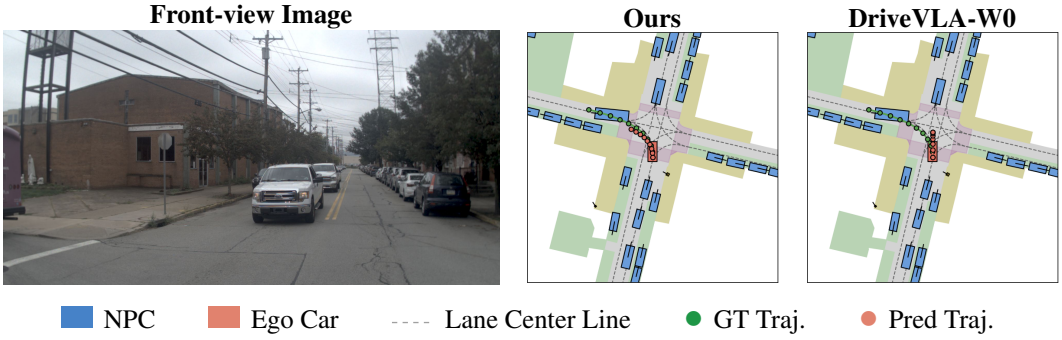



Figure 6: **Left turn in a small intersection with the limited view.** DriveVLA-W0 baseline goes straight at a small intersection where it is supposed to turn left. Our method turns left even though the front view is limited.

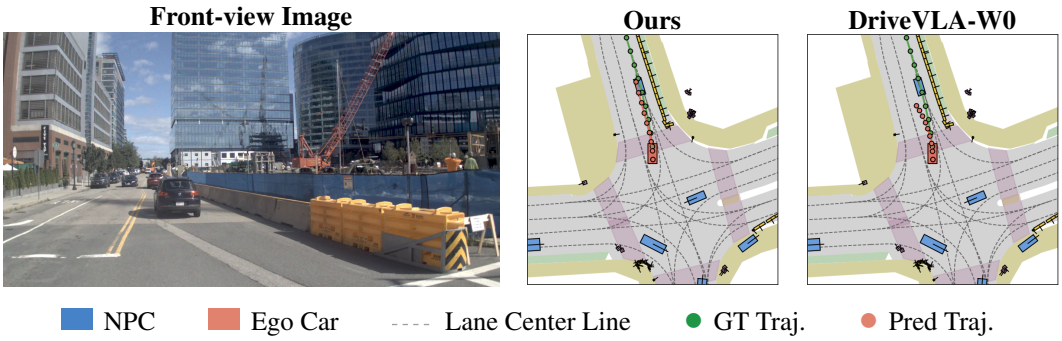



Figure 7: **Turn left at the construction intersection.** DriveVLA-W0 baseline goes into the opposite lane at the construction intersection. Our method turns left and keep in the right lane.

### C.3 FAILURE CASE

To provide an unbiased account of our model's performance and delineate paths for subsequent enhancement, representative failure cases encountered in evaluation are analyzed. They fall into two main types: deficits in capturing critical small elements, and constraints of front-view cameras under turning scenarios, as shown in Figure 8, Figure 9 and Figure 10.

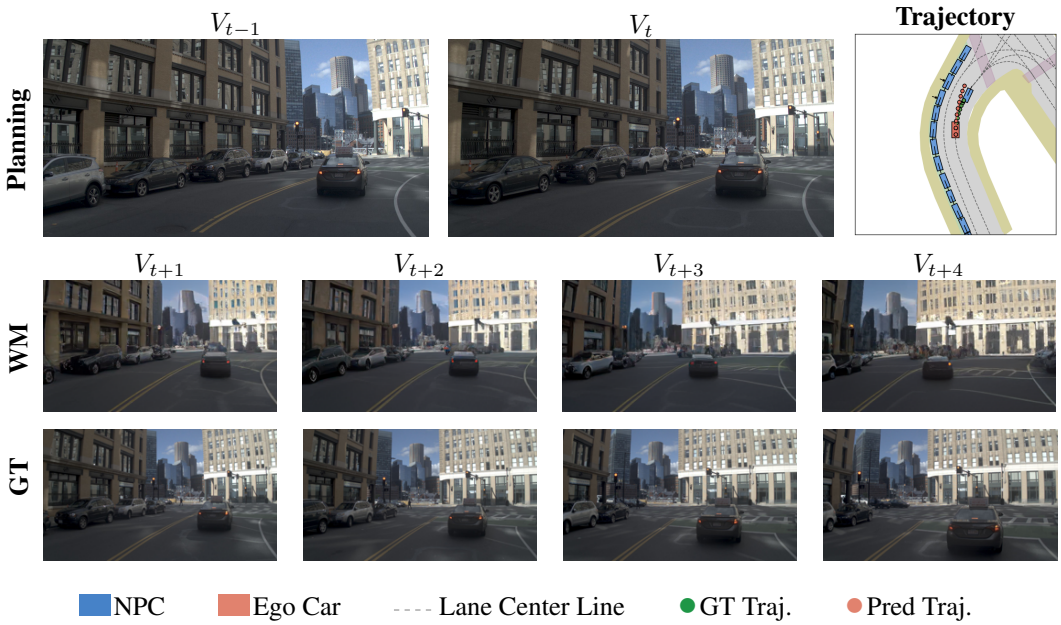



Figure 8: **Failure case analysis: deficits in capturing critical small elements.** We illustrate a failure case in which our model attempts to nudge a queuing vehicle ahead. The world model's predictions reveal that the model misclassifies the queuing vehicle as low-speed moving and tries to nudge past it for higher traffic efficiency, failing to recognize the queuing state indicated by the red light ahead and the vehicle's brake lights. Since this critical cue is only discernible from a few pixels, fine-grained element capture and causal reasoning remain a core future research direction.

### C.4 MORE SCENE NOISE ROBUSTNESS CASE

To provide a balanced view of our model's scene noise robustness, we perform qualitative analyses on illumination interference and blur noise cases. Our model exhibits more stable and reasonable performance under noise than the baseline. Besides the left-turn scenario in the main text, we compare output variations in typical right-turn and straight-ahead scenarios by adjusting illumination intensity alone, verifying that our model yields more consistent outputs under illumination perturbations, as shown in Figure 11a and Figure 11b. We further present cases with natural non-uniform blur from rain and fog, where traffic lights and vehicles are partially or fully occluded. The pure pixel-regression baseline shows severe degradation: it crawls during red-light queuing and acts overly conservative to slow down. By contrast, benefiting from the LWM's inherent scene noise robustness, our model maintains accurate and reasonable driving behaviors under the same noisy inputs, with trajectories more aligned with human driving patterns, as shown in Figure 12 and Figure 13.

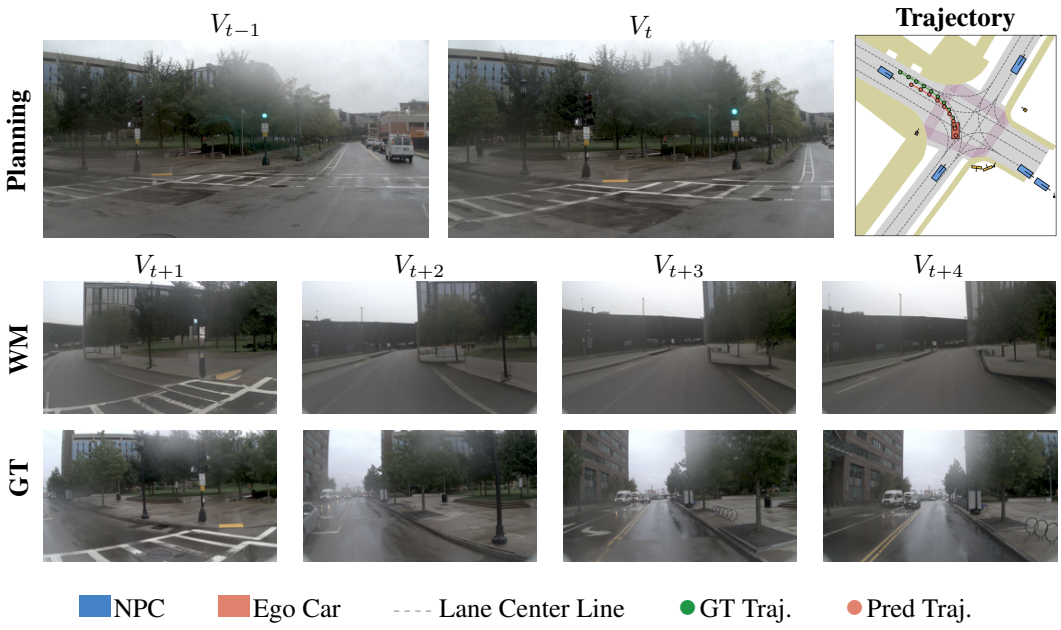
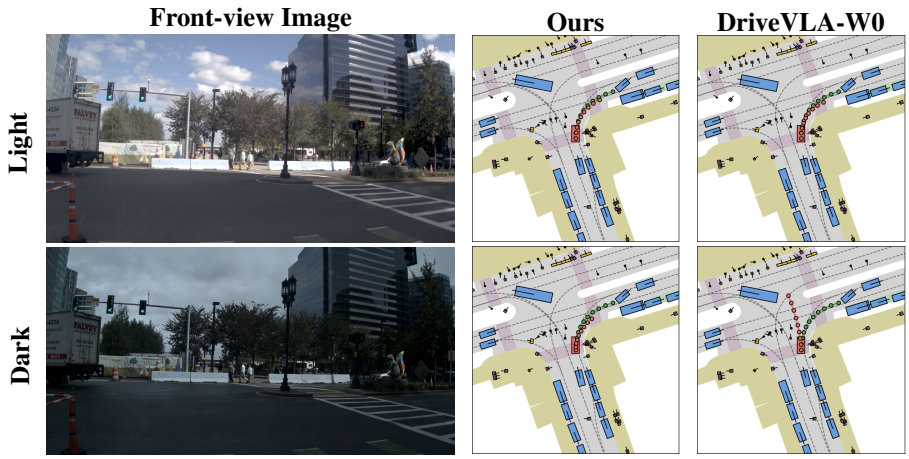


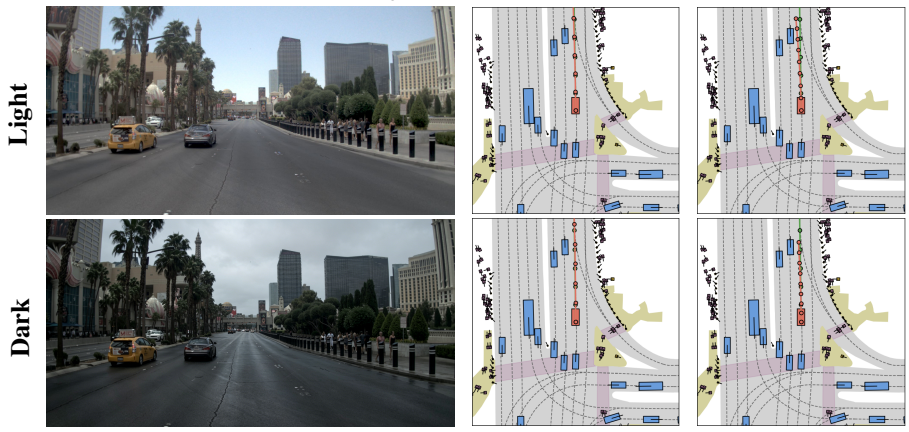

Figure 9: **Failure case analysis: constraints of front-view cameras under left turn.** We present a failure case in which our model crosses the central double yellow line during left turns. Limited by field of view, historical and current frames fail to deliver valid cues for the target lane in the left-turn process; the world model's predictions correspondingly lack the double road marking. This demonstrates the importance of multi-view camera inputs for turning scenarios, while trading off camera count against inference latency remains a key unresolved issue.




Figure 10: **Failure case analysis: constraints of front-view cameras under right turn.** We present a failure case in which our model rear-ends a queuing leading vehicle during right turns. Limited by field of view, valid cues of surrounding vehicles are unavailable in the right-turn process, and the world model's predictions correspondingly miss the preceding vehicle. This again demonstrates the necessity of multi-view camera inputs for turning scenarios.




(a) Turn right in the intersection.




(b) Go straight in the intersection.




Figure 11: **Turn right and go straight under the fluctuating illumination.** (a) DriveVLA-W0 goes straight under right turn instruction, while our method remains robust. (b) DriveVLA-W0's behavior changes from changing lanes to going straight, while our approach remains going straight.

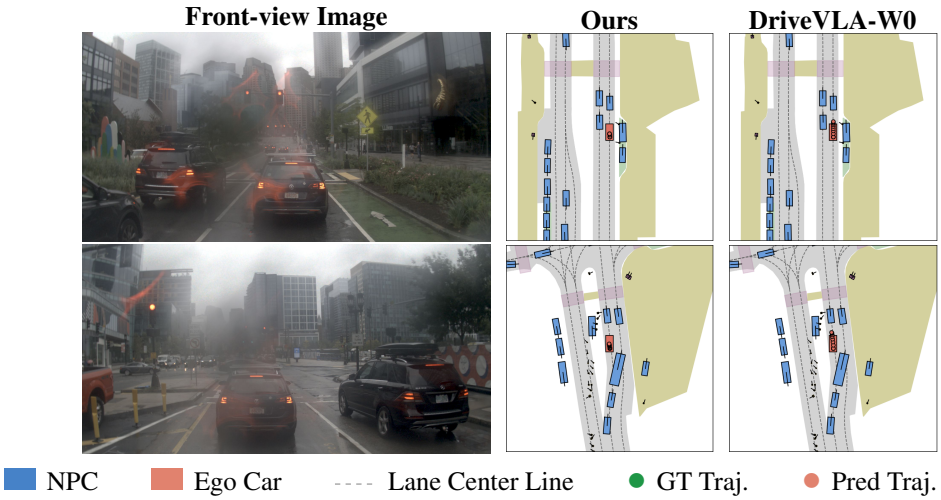



Figure 12: **Remain stationary at red-light intersection under the blur noise.** DriveVLA-W0 baseline crawls during red-light queuing, while our method remains stationary.




Figure 13: **Go straight under the blur noise.** DriveVLA-W0 baseline try to slow down due to the blurry image. Our method maintains a reasonable speed and is more aligned with human driving patterns.