# Learning Vision-Language-Action World Models for Autonomous Driving

## Supplementary Material

In the supplementary material, we first present the methodology detail of our proposed VLA-World, including Group Relative Policy Optimization (GRPO), short-term trajectory prediction, and theoretical analysis of VLA-World. Then, we provide the details of the datasets and implementation. Furthermore, we present additional experimental results to demonstrate the effectiveness of VLA-World.

## A. Method Details

### A.1. Group Relative Policy Optimization

In the final training stage of VLA-World, we adopt Group Relative Policy Optimization (GRPO) [9] to unleash the latent reasoning and decision-making capabilities. Unlike traditional PPO [8], which relies on a computationally heavy value function (Critic) that often struggles with high-dimensional visual dynamics, GRPO operates in a value-free paradigm. It leverages group-wise statistics to estimate baselines, significantly reducing memory overhead while stabilizing training.

For each driving scenario prompt $o$, the current policy $\pi_\theta$ samples a group of $G$ candidate rollouts (outputs), denoted as $\{o, o_1, \ldots, o_G\}$. These candidates represent diverse reasoning paths, ranging from conservative yielding to assertive maneuvering. Instead of relying on a neural reward model, we employ a set of lightweight, rule-based verifiers to compute rewards. These include outcome rewards (e.g., collision checking, generation quality, temporal consistency) and format rewards (e.g., strict compliance with the required output structure). Each rollout is evaluated to produce a scalar reward set $\{r_1, r_2, \ldots, r_G\}$.

To determine the relative quality of each reasoning path, we compute the normalized advantage for the $i$-th rollout within the group:

$$A_i = \frac{r_i - \mu}{\sigma}, \quad \mu = \frac{1}{G}\sum_j r_j,\ \sigma = \mathrm{std}(r_1, \ldots, r_G) \quad (1)$$

This group-based normalization effectively serves as a dynamic baseline, encouraging the model to prioritize trajectories that outperform the group average. The policy is then updated by maximizing the following surrogate objective:

$$J(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\min\left(\frac{\pi_\theta(\tau_i \mid o)}{\pi_{\theta_{\text{old}}}(\tau_i \mid o)}A_i, \text{clip}\right)\right] - \beta D_{\text{KL}}(\pi_\theta, \pi_{\text{old}}). \quad (2)$$

where the KL-divergence term ensures the policy does not deviate excessively from the reference model (the SFT checkpoint), preventing reward hacking.

By optimizing this objective, VLA-World effectively performs *Self-Verification*: it learns to implicitly discard hallucinatory or unsafe trajectories and reinforces the internal *chain-of-thought* that leads to compliant and safe driving behaviors. This mechanism allows the model to refine its logical consistency purely through rule-based feedback, resulting in a robust planner that is both explainable and physically grounded.

### A.2. Short-term Trajectory Prediction

To ensure the synthesized future views are physically plausible and consistent with the vehicle's movement, we employ a physics-grounded trajectory predictor. This module estimates the ego-vehicle's future position $\hat{\mathbf{P}}_{t+\tau}$ at a look-ahead horizon $\tau$ (e.g., 0.5s), conditioned on both the historical state sequence $\mathcal{H}$ and the high-level mission goal $g$ (e.g., *Left*). We formulate this prediction as a superposition of inertial dynamics and intentional control.

**Kinematic State Estimation.** First, we extract the vehicle's instantaneous kinematic state from the discrete historical trajectory $\mathcal{H} = \{\mathbf{P}_{t-N}, \ldots, \mathbf{P}_t\}$, where $\mathbf{P}_i \in \mathbb{R}^2$ denotes the coordinates in the ego-frame. We approximate the current velocity $\mathbf{v}_t$ and the historical inertial acceleration $\mathbf{a}_{\text{hist}}$ using a finite difference method:

$$\mathbf{v}_t = \frac{\mathbf{P}_t - \mathbf{P}_{t-1}}{\Delta t}, \quad \mathbf{a}_{\text{hist}} = \frac{\mathbf{v}_t - \mathbf{v}_{t-1}}{\Delta t} \quad (3)$$

where $\Delta t$ represents the sampling interval. The term $\mathbf{a}_{\text{hist}}$ captures the vehicle's momentum prior to any new control inputs.

**Intention-Driven Refinement.** A pure constant-acceleration model often fails to capture sudden maneuvers dictated by the mission goal. To address this, we introduce a goal-conditioned acceleration term $\mathbf{a}_{\text{goal}}$. The navigational command $c$ is mapped to a target spatial offset or a virtual waypoint, implying a required trajectory deviation. We derive $\mathbf{a}_{\text{goal}}$ as the constant acceleration required to shift the vehicle from its current state $\mathcal{S}_t = \{\mathbf{P}_t, \mathbf{v}_t\}$ to the target state determined by $c$ within the horizon $\tau$.

**Fusion and Prediction.** The final predicted trajectory is modeled as a linear fusion of the historical inertia and the future intention. We define the effective acceleration $\mathbf{a}_{\text{eff}}$ using an adaptive weighting factor $\lambda \in [0, 1]$:

$$\mathbf{a}_{\text{eff}} = (1 - \lambda)\mathbf{a}_{\text{hist}} + \lambda\mathbf{a}_{\text{goal}} \quad (4)$$

where $\mathbf{a}_{\text{goal}} = \frac{2}{\tau^2}(\Delta\mathbf{P}_{\text{ideal}} - \mathbf{v}_t\tau)$, and $\Delta\mathbf{P}_{\text{ideal}}$ represents the theoretical displacement required by the command $c$. Consequently, the predicted position $\hat{\mathbf{P}}_{t+\tau}$ is computed via

the kinematic equation:

$$\hat{\mathbf{P}}_{t+\tau} = \mathbf{P}_t + \mathbf{v}_t\tau + \frac{1}{2}\mathbf{a}_{\text{eff}}\tau^2 \tag{5}$$

This formulation allows our model to seamlessly transition between momentum-based continuity (e.g., straight-line driving) and intention-based maneuvering (e.g., sharp turns), providing a robust geometric prior for the subsequent frame generation process.

### A.3. Theoretical Analysis of VLA-World

In this section, we provide a theoretical analysis for VLA-World by formalizing autonomous driving as a joint optimization problem. We demonstrate that VLA-World aligns better with the driving objective than independent VLA or World Model paradigms.

**The Joint Modeling Objective.** The central object of autonomous driving is the joint distribution of the planned ego-trajectory $\tau_{t:t+H}$ and the anticipated short-term future environment $x_{t+1}$, conditioned on observation history $o_{1:t}$ and goal $g$. According to the probability chain rule, this joint distribution factorizes as:

$$p(\tau_{t:t+H}, x_{t+1} \mid o_{1:t}, g) = \underbrace{p(\tau_{t:t+H} \mid o_{1:t}, g)}_{\text{Policy (Decision)}} \cdot \underbrace{p(x_{t+1} \mid o_{1:t}, \tau_{t+1})}_{\text{World Model (Imagination)}} \tag{6}$$

We define the ultimate driving objective $J(\omega)$ as the expected task return $R$ (aggregating safety, comfort, and rule compliance) over this joint distribution:

$$J(\omega) = \mathbb{E}_{p_\omega(\tau, x \mid o, g)}\big[R(\tau_{t:t+H}, x_{t+1})\big] \tag{7}$$

Learning to drive is thus equivalent to learning a parameter set $\omega$ that shapes this joint distribution to maximize reward. VLA-World explicitly parameterizes and optimizes both factors in Eq. (6), whereas previous paradigms only address one.

**Analysis of VLA Models.** A pure VLA model implicitly integrates out the future state $x_{t+1}$, modeling only the marginal policy distribution:

$$\pi_{\text{VLA}}(\tau \mid o, g) \approx \int p^\star(\tau, x \mid o, g)\, dx \tag{8}$$

From a variational inference perspective, ignoring the explicit future state $x$ leads to a loose approximation of the optimal policy. For any auxiliary distribution $q(x \mid o, \tau)$ describing the environment dynamics, the log-likelihood of the optimal policy is bounded by the Evidence Lower Bound (ELBO):

$$\log p^\star(\tau \mid o, g) \geq \mathbb{E}_{x\sim q}\big[\log p^\star(\tau, x \mid o, g) - \log q(x \mid o, \tau)\big] \tag{9}$$

Theoretical Insight: A VLA model that discards $x_{t+1}$ is mathematically equivalent to optimizing a loose lower bound where the predictive information about scene evolution is lost. It tries to match the marginal directly without understanding the underlying causal variable $x$. In contrast, VLA-World models the joint numerator $p(\tau, x \mid o, g)$ directly. By explicitly generating $x_{t+1}$, VLA-World tightens this bound, effectively using the "imagined" future to reduce the uncertainty in policy estimation.

**Analysis of World Models.** Classical world models focus on learning the transition dynamics $p_{\text{WM}}(x_{t+1} \mid o, \tau)$ via a reconstruction objective:

$$J_{\text{WM}}(\theta) = \mathbb{E}\big[-\log p_\theta(x_{t+1} \mid o, \tau)\big] \tag{10}$$

Crucially, this objective is weakly coupled to the driving decision. A world model seeks to maximize pixel fidelity, not driving safety. The planning is typically performed by a separate search procedure on top of this frozen model:

$$\tau^{\text{WM}} = \arg\max_\tau \mathbb{E}_{x\sim p_{\text{WM}}}[R(\tau, x)] \tag{11}$$

Theoretical Insight: Any mismatch between generative accuracy (reconstruction) and planning utility (safety) creates a performance bottleneck. A high-fidelity simulation of a collision is valid for Eq. (11) but disastrous for the agent. Unlike VLA-World, pure world models do not back-propagate the decision reward $R$ into the model parameters $\theta$, leaving the *imagination* disconnected from the *consequence*.

**Analysis of VLA-World.** VLA-World unifies the policy $\pi_\omega$ and world model $p_\omega$ into a single autoregressive transformer. The gradient of our objective $J(\omega)$ (from Eq. (7)) naturally decomposes to update both components:

$$\nabla_\omega J(\omega) = \mathbb{E}\Big[\underbrace{\nabla_\omega \log \pi_\omega(\tau \mid o, g) \cdot R}_{\text{Policy Gradient}} + \underbrace{\nabla_\omega \log p_\omega(x \mid o, \tau) \cdot R}_{\text{World Model Gradient}}\Big] \tag{12}$$

This reveals the core mechanism: Both the decision term and the imagination term are optimized by the same driving reward $R$. In our implementation, we employ reinforcement learning with GRPO. Let $u$ denote the full sequence of tokens including trajectory $\tau$, future frame $x$, and reasoning language. We maximize:

$$\mathcal{J}_{\text{GRPO}}(\omega) = \mathbb{E}_{u\sim\pi_\omega}\big[\log \pi_\omega(u \mid o, g) \cdot A(u)\big] \tag{13}$$

Because $u$ contains the future-image tokens, the *imagination* is no longer just minimizing reconstruction error; it is being reinforced to generate futures that lead to high-reward outcomes (e.g., highlighting risks that aid safety). This forms an imagination-decision loop. Finally, we show that VLA-World is a strictly more expressive hypothesis class than either baseline.

**VLA as a special case:** If we mask the imagination branch (force $p_\omega(x \mid \dots)$ to be a delta function or ignore it), Eq. (6) collapses to the marginal policy $\pi_\omega(\tau \mid o, g)$, recovering a standard VLA.

**World Model as a special case:** If we freeze the parameters of $p_\omega(x \mid o, \tau)$ and use an external optimizer for $\tau$, we recover the trajectory search of standard World Models (Eq. (11)).

## B. Experiments

### B.1. Dataset

We evaluate trajectory planning and future frame generation on the nuScenes dataset [1] following the traditional end-to-end methods [2, 4, 6], VLA [5, 12, 15] and world models [7, 13, 14, 16]. The nuScenes comprises 1,000 driving scenes, each about 20 seconds long, recorded with a 32-beam LiDAR and six cameras offering a full 360-degree view. The dataset includes 28,130 training samples, 6,019 validation samples, and 193,082 unlabeled samples.

**Pretraining Stage.** To endow the VLM with an intuitive understanding of physical dynamics, we construct a visual generation pretraining dataset as shown in Fig. 1 (a) (≈500k). In this stage, the model functions strictly as a generative world model. The input prompt consists of current multi-view observations alongside explicit definitions of the ego-centric coordinate system and physical units. The objective is to autoregressively predict the discrete visual tokens corresponding to a future frame (e.g., $\Delta t = 0.5s$) for a specified camera view. This pretraining forces the model to internalize spatiotemporal evolution laws, such as agent motion and ego motion from large-scale data, establishing a foundational *imagination* capability without the complexity of high-level linguistic reasoning.

**SFT and RL Stages.** For the Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL) stages, we introduce a multi-step learning paradigm as illustrated in Fig. 1 (b) (≈20k). The input is augmented with detailed vehicle kinematics (velocity, acceleration), historical trajectories, and high-level mission commands. The model output is structured into a causal reasoning sequence: it first parses the scene via <perception> and estimates a short-term <prediction>, which conditions the generation of the future <visual> frame. Crucially, the model then explicitly reasons over this imagined future in the <think> block to assess potential risks before determining the high-level <action> and regressing the precise long-term trajectory points in <answer>. This structure unifies generation and planning, allowing GRPO-based RL to optimize the consistency between the imagined future and the executed safety maneuvers.

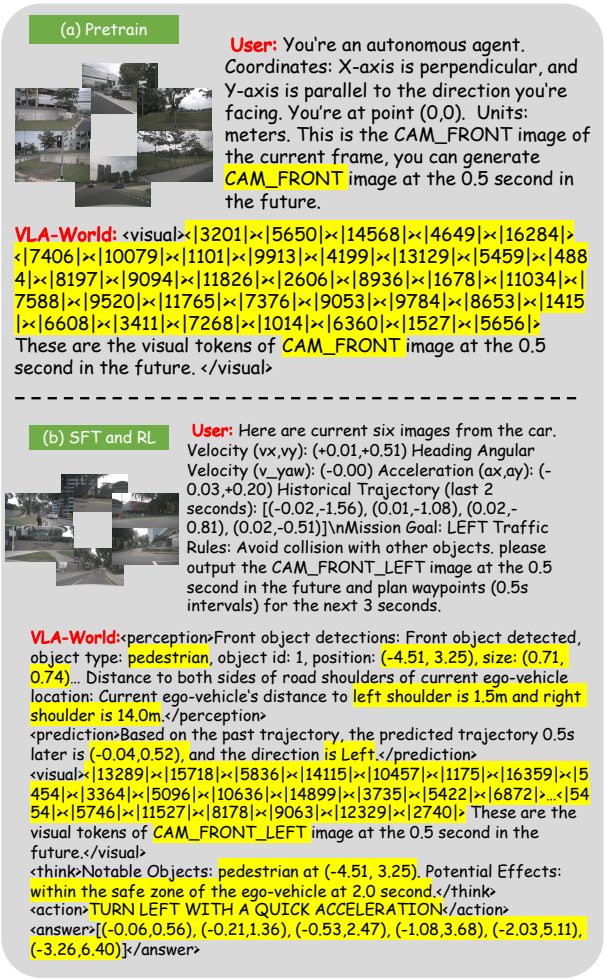


Figure 1. Data sample of (a) pretraining stage, (b) supervised fine-tuning, and (c) reinforcement learning.

### B.2. Implementation Details

We build our model upon the Qwen-VL family [11]. All training stages, including pretraining, supervised fine-tuning, and reinforcement learning, are conducted on 8 A100 GPUs, and inference is performed on 4 A100 GPUs. The pretraining and supervised fine-tuning stages use the LLaMA Factory framework [17], and the reinforcement learning stage is trained with the Easy-R1 framework [10]. We adopt multi-view images as input and set the maximum pixel count to 524,288, with a gradient accumulation step of 2. During pretraining, the model is trained for 30 epochs using AdamW with an initial learning rate of $5 \times 10^{-4}$ and a per-device batch size of 16. For supervised fine-tuning, we train for 12 epochs with AdamW and an initial learning rate of $1 \times 10^{-4}$. Starting from the supervised fine-tuning checkpoint, we further optimize the model for one epoch using Group Relative Policy Optimization. The policy is trained with a learning rate of $1 \times 10^{-6}$ and a global batch size of 16. To retain the


Figure 2. Comparison between our VLA-World and the state-of-the-art FSDrive [15] on generating the future frame at next 0.5 seconds.

Table 1. Evaluation of trajectory planning L2 errors (ST-P3) on nuScenes with varying input view resolutions.

| Res. | L2 Error (m) ↓ | | | |
|---|---|---|---|---|
| | 1s | 2s | 3s | Avg. |
| 36000 | 0.03 | 0.14 | 0.98 | 0.38 |
| 52884 | 0.11 | 0.27 | 0.52 | 0.30 |

Table 2. Evaluation of trajectory planning L2 errors (ST-P3) on nuScenes with varying model sizes.

| Method | L2 Error (m) ↓ | | | |
|---|---|---|---|---|
| | 1s | 2s | 3s | Avg. |
| Qwen2-VL-2B | 0.11 | 0.27 | 0.52 | 0.30 |
| Qwen2.5-VL-3B | 0.05 | 0.08 | 0.76 | 0.29 |
| Qwen2-VL-7B | 0.03 | 0.03 | 0.47 | 0.18 |

Table 3. Evaluation of trajectory planning L2 errors (ST-P3) on nuScenes with training strategy.

| Method | L2 Error (m) ↓ | | | |
|---|---|---|---|---|
| | 1s | 2s | 3s | Avg. |
| w/o. Mixed | 0.27 | 0.47 | 0.73 | 0.49 |
| Qwen2-VL-2B | 0.11 | 0.27 | 0.52 | 0.30 |

behavior learned during supervised fine-tuning and ensure stable optimization, we apply a KL divergence regularization term with a coefficient of $1 \times 10^{-2}$. For each prompt, we sample 8 candidate responses to estimate the policy gradient. A cosine learning rate scheduler with a warm-up ratio of 0.1 is applied throughout all training stages to stabilize early optimization. We evaluate trajectory planning performance using L2 displacement error and collision rate, following widely adopted protocols in prior work [2, 3, 6, 12, 15]. UniAD [4] computes both metrics at each individual timestep, whereas ST P3 [3] and VAD [2, 6] report the average values over all preceding timesteps. For fair comparison, we follow the respective evaluation strategies of each method. In addition, consistent with recent approaches in generative prediction [13, 14], we adopt the Fréchet Inception Distance to quantify the visual fidelity of synthesized future frames.

### B.3. More Discussion

**Effectiveness of Input Resolution.** Tab. 1 investigates the sensitivity of our model to input view resolutions. The results demonstrate that higher resolution inputs generally yield better planning performance, particularly over longer time horizons. Although the lower resolution $(36,000)$ is competitive at short intervals $(1s)$, the higher resolution model $(52,884)$ demonstrates superior robustness, achieving the lowest average L2 error of $0.30$m. This indicates that maintaining high-fidelity visual information is crucial for mitigating error accumulation in trajectory prediction.

**Effectiveness of Model Size.** Tab. 2 presents an ablation study on the effect of model size by varying the backbone among Qwen2-VL-2B, Qwen2.5-VL-3B, and Qwen2-VL-7B. The results demonstrate a clear scaling law: increasing the model capacity significantly enhances trajectory planning performance. The Qwen2-VL-7B model achieves state-of-the-art results with an average L2 error of 0.18m, outperforming the 2B and 3B variants by a substantial margin (approximately $40\%$ relative improvement). This suggests that the stronger reasoning and generalization capabilities inherent in larger parameters are essential for handling the complex causal dependencies in autonomous driving scenarios, particularly for maintaining accuracy over longer time horizons (e.g., reducing 3s error to 0.47m).

**Effectiveness of Training Strategy.** In Tab. 3, we conduct an ablation study to verify the contribution of our multi-task

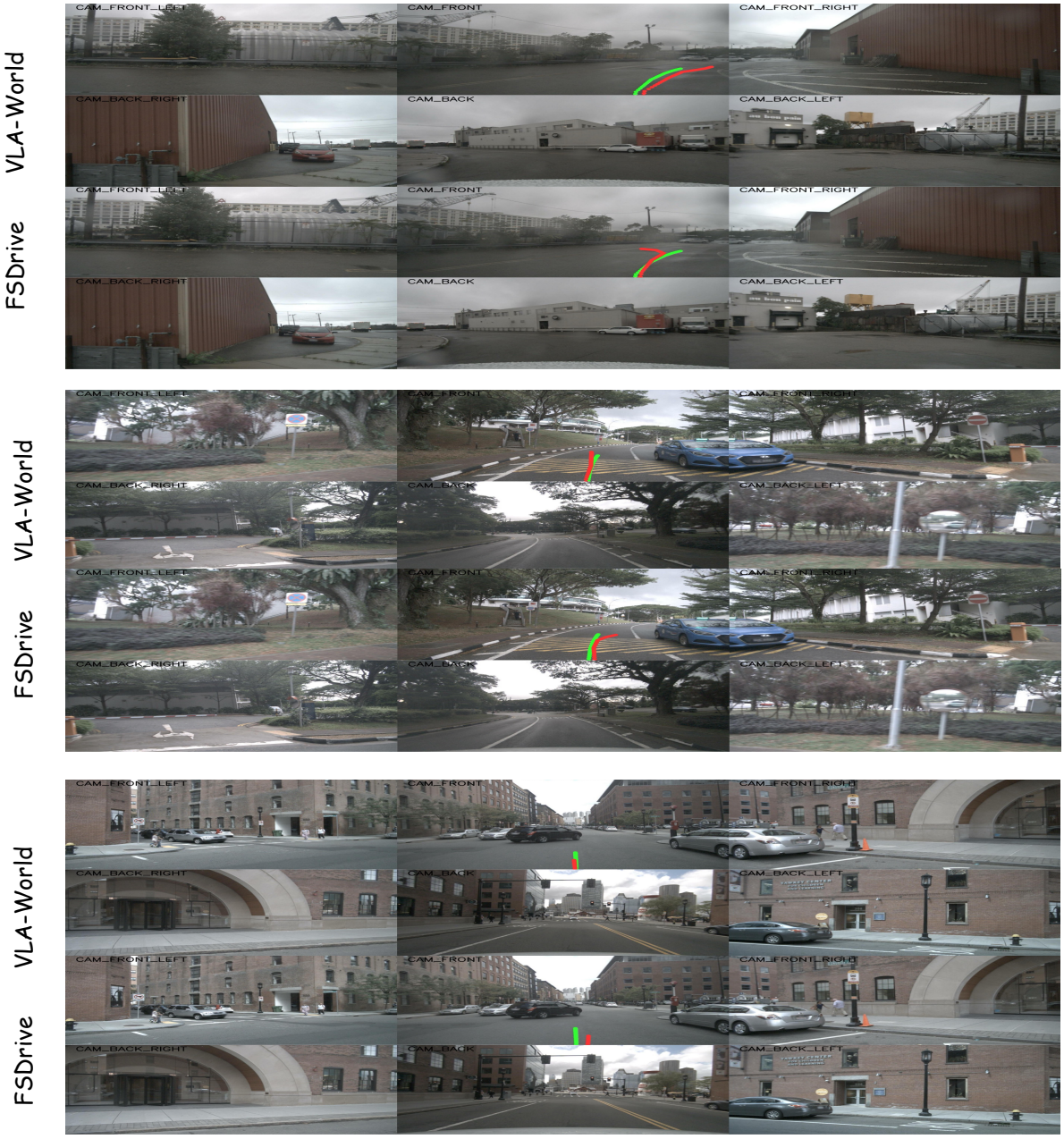


Figure 3. Comparison of 3-second future trajectory predictions generated by our VLA-World and the state-of-the-art FSDrive [15]. Zoom in for a better view.

mixed dataset. We compare the full Qwen2-VL-2B model against a baseline trained without mixed data (*w/o. Mixed*). The results reveal that removing the diverse supervision signals leads to a significant performance degradation, with the average L2 error increasing from 0.30m to 0.49m. This substantial gap underscores the critical role of mixed-task training (combining perception, reasoning, and planning) in learning robust feature representations, enabling the model to generalize better across varying time horizons.

### B.4. More Visualization

We visualize the generation and trajectory planning results with the SOTA FSDrive [15]. More visualization results can be found in the video demo in our supplementary materials.

**Generation Results.** We present a qualitative comparison of the generated 0.5s future frames in Fig. 2. As illustrated by the red-highlighted regions, the baseline method (FSDrive, bottom row) struggles to maintain object coherence during the prediction horizon. It exhibits noticeable artifacts, including geometric distortion of vehicles and a loss of high-frequency details in the background, indicating a lack of robust spatiotemporal constraints. Conversely, VLA-World (top row) demonstrates significantly improved visual fidelity. By effectively leveraging the trajectory-aware conditioning, our model preserves the structural rigidity of dynamic agents and the sharpness of the scene. The generated frames exhibit high photorealism and consistency, validating that our short-term prediction successfully mitigates the *hallucination* artifacts common in pure VLA.

**Trajectory Planning.** We provide visualizations of the 3-second future trajectories predicted by VLA-World and FSDrive. As shown in Fig. 3, VLA-World produces noticeably more precise trajectory predictions than FSDrive, especially near the 3-second horizon where the deviation from the ground truth becomes minimal. This improvement stems from our paradigm: first predicting the future state, then generating the corresponding 0.5-second future frames based on that prediction, and finally reasoning over the imagined scene to refine the outcome. In contrast, FSDrive lacks such reflective and iterative reasoning capabilities, which leads to cumulative drift over longer temporal horizons.

## References


[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11621–11631, 2020. 3

[2] Shaoyu Chen, Bo Jiang, Hao Gao, Bencheng Liao, Qing Xu, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Vadv2: End-to-end vectorized autonomous driving via probabilistic planning, 2024. 3, 4

[3] Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *ECCV*, pages 533–549, 2022. 4

[4] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *CVPR*, pages 17853–17862, 2023. 3, 4

[5] 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. 3

[6] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *ICCV*, pages 8306–8316, 2023. 3, 4

[7] Seung Wook Kim, Jonah Philion, Antonio Torralba, and Sanja Fidler. Drivegan: Towards a controllable high-quality neural simulation. In *CVPR*, pages 5820–5829, 2021. 3

[8] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 1

[9] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 1

[10] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024. 3

[11] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3

[12] Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In *CVPR*, pages 22442–22452, 2025. 3, 4

[13] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv:2309.09777*, 2023. 3, 4

[14] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *CVPR*, pages 14749–14759, 2024. 3, 4

[15] 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. *arXiv preprint arXiv:2505.17685*, 2025. 3, 4, 5

[16] 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. 3

[17] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *ACL*, 2024. 3