PHYSDRIVE: A PHYSICALLY GROUNDED VISION-LANGUAGE MODEL FOR DRIVING UNDERSTANDING AND PLANNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language-Action (VLA) models have emerged as a promising paradigm for autonomous driving, supporting both scene understanding and motion planning. However, existing VLA models are typically optimized primarily with downstream outputs such as language responses or expert trajectories, providing limited supervision for the intermediate representations that connect visual perception to decision making. This output-centric training may encourage shortcut learning and hinder the model from capturing the underlying 3D geometry and physical dynamics of driving scenes. To address this limitation, we propose **PhysDrive**, a physically grounded vision-language model that uses 4D occupancy as structured training-time supervision for its internal representations. Specifically, PhysDrive decomposes 4D occupancy supervision into two complementary components: the current occupancy grounds the model in the 3D geometry of the observed scene, while future occupancy guides the modeling of its physical evolution over time. Together, these signals encourage PhysDrive to encode spatial structure and scene dynamics before producing downstream understanding or planning outputs. We further introduce a reinforcement learning framework to align generated trajectories with driving-oriented objectives, including safety, drivable-area compliance, and progress. Extensive experiments on NuInstruct, nuScenes, and NAVSIM demonstrate that PhysDrive consistently improves driving understanding and planning, demonstrating the effectiveness of physically grounded representations for downstream driving tasks.

1 INTRODUCTION

Vision-Language Models (VLMs) have been increasingly adopted for driving scene understanding and planning owing to their strong generalization capabilities (Xu et al. 2024; Tian et al. 2024). Follow-up works further extend VLMs into Vision-Language-Action (VLA) models by incorporating dedicated action experts for planning (Fu et al. 2025; Li et al. 2026d). Existing VLM- and VLA-based approaches typically take visual observations of driving scenes and textual instructions as input and generate textual responses or waypoints. However, most of these models are only supervised by expert responses or trajectories, which may lead to shortcut learning rather than a proper physical understanding of the 3D spatial structure and temporal evolution of driving scenes.

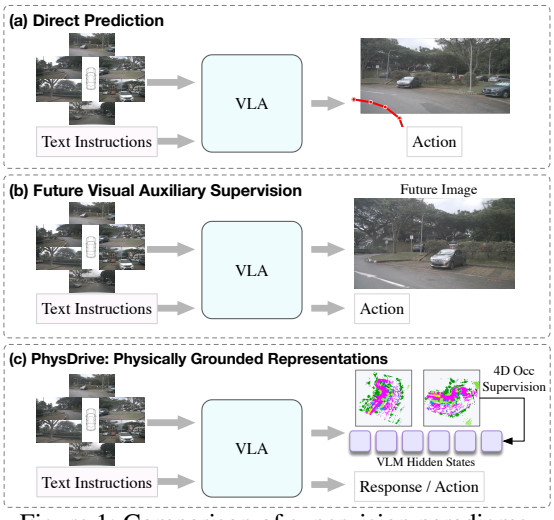

Figure 1: Comparison of supervision paradigms.

To provide richer supervision beyond final responses or trajectories, recent works have explored future-scene prediction as an auxiliary learning signal. A common strategy is to predict future visual observations before or alongside downstream

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.

planning. For example, FSDrive (Zeng et al. 2026) predicts future frames augmented with lane dividers and 3D bounding boxes, while ImagiDrive (Li et al. 2025a), DriveVLA-W0 (Li et al. 2026c), and OneVL (Lu et al. 2026) similarly exploit future visual prediction to encourage the model to capture how driving scenes evolve over time. Such future visual supervision provides richer temporal cues than direct output supervision alone, but it mainly encourages the model to match appearance changes in the 2D image space, without explicitly constraining the underlying 3D spatial structure of the driving scene. X-Mind further explores future prediction in the Bird's-Eye-View (BEV) space to better capture scene-level dynamics, yet BEV representations inevitably compress the physical world onto the ground plane. These limitations motivate a more direct form of supervision that can ground the model's internal representations in both the 3D geometry of the current scene and its physical evolution over time.

A natural choice for such physical grounding is 4D occupancy, which explicitly represents both the 3D geometry of the current scene and its temporal evolution over time. Existing occupancy-centric methods have demonstrated the value of such representations for autonomous driving. OccWorld jointly performs occupancy forecasting and planning (Zheng et al. 2024), while subsequent methods employ large language models to unify occupancy modeling and action generation (Wei et al. 2024; Xu et al. 2025). However, directly using occupancy as model input requires an additional perception pipeline and may discard fine-grained visual cues, such as traffic-light states, that are naturally preserved in image observations. A more flexible alternative is therefore to retain images as the model input while using occupancy as structured supervision for representation learning. OccVLA adopts multi-view images as input and aligns VLM representations with current occupancy (Liu et al. 2025), providing explicit 3D geometric supervision but without modeling future scene evolution. This limitation is particularly important for autonomous driving, where understanding the dynamic evolution of the scene is essential for capturing physical interactions and supporting downstream planning.

To address these limitations, we propose **PhysDrive**, a physically grounded vision-language model for driving understanding and planning. PhysDrive uses structured 4D physical supervision during training to shape the internal representations of the VLM. Specifically, current occupancy grounds the model in the 3D geometry of the observed scene, while future occupancy guides the modeling of its physical dynamics over time. For driving-understanding tasks, we further introduce region-of-interest (RoI) supervision to encourage task-relevant spatial grounding. Together, these complementary signals enable PhysDrive to learn physically grounded representations that capture scene geometry, dynamic evolution, and question-relevant spatial information for downstream understanding and planning. We further introduce a reinforcement learning framework tailored to PhysDrive to align generated trajectories with driving-oriented objectives such as collision avoidance, drivable-area compliance, and driving progress.

Our contributions are summarized as follows:

- We propose **PhysDrive**, a physically grounded vision-language model that uses structured physical supervision to improve both driving understanding and planning.
- We introduce complementary grounding signals from 4D occupancy to capture 3D scene geometry and physical dynamics, together with optional RoI supervision for question-relevant region grounding.
- We develop a reinforcement learning framework to further align PhysDrive's planning outputs with driving-oriented objectives.
- Extensive experiments on NuInstruct, nuScenes, and NAVSIM demonstrate that PhysDrive consistently improves driving understanding and planning.

2 Related Work

Driving VLMs and VLAs. Existing driving VLMs and VLAs, such as DriveGPT4, OmniDrive, OpenDriveVLA, and AutoVLA, typically encode multi-view images, ego states, and navigation commands with a pretrained vision-language model and optimize the model to produce downstream outputs such as textual responses, waypoints, or control signals (Xu et al. 2024; Wang et al. 2025; Zhou et al. 2026a;b). Follow-up works, including ORION, RecogDrive, Drive-PI, and SpaceDrive, introduce stronger action representations or dedicated action heads to better connect VLM features

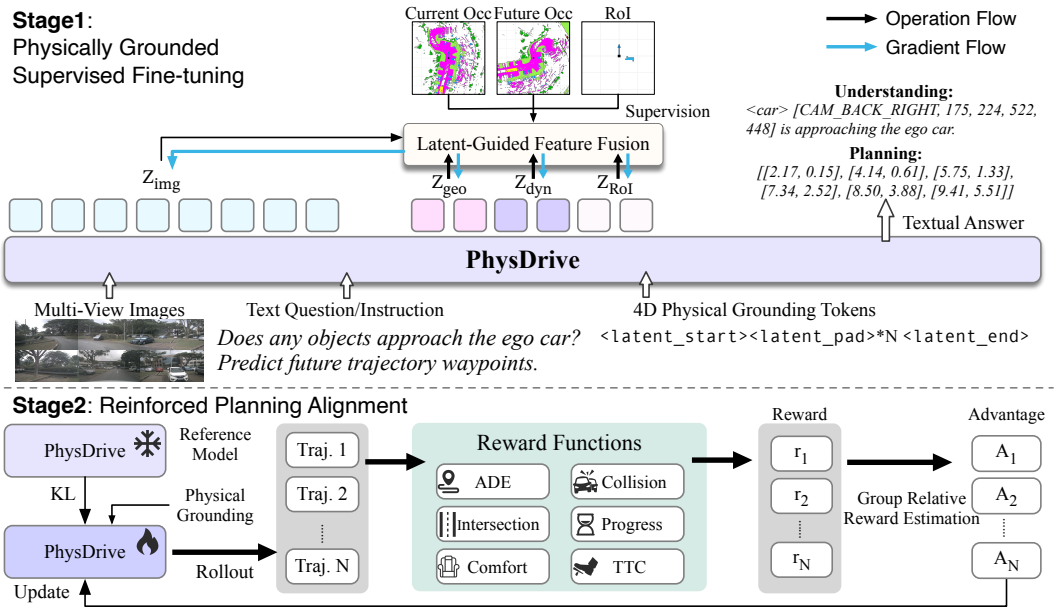

Figure 2: **Overview of PhysDrive.** Stage 1 learns physically grounded representations with 4D occupancy and optional RoI supervision. Stage 2 applies GRPO to align planning outputs with driving-oriented rewards while preserving the learned 4D representation.

with continuous planning outputs (Li et al. 2026d; Liu et al. 2026; Fu et al. 2025; Li et al. 2026b). Despite their strong performance, many of these methods are primarily optimized with task-level supervision, such as language responses or expert trajectories. Such supervision provides limited constraints on the internal representations that bridge visual perception and downstream decision making, making it difficult to explicitly enforce the encoding of 3D scene geometry and dynamic scene evolution. In contrast, PhysDrive introduces structured physical supervision to shape the internal representations of the VLM, enabling physically grounded representations that support both driving understanding and planning.

Future Visual Auxiliary Supervision. Recent works introduce future-scene prediction as an additional supervision signal beyond direct task outputs. FSDrive predicts future frames augmented with lane dividers and projected 3D boxes (Zeng et al. 2026), while ImagiDrive, Epona, DriveLaW, DriveVLA-W0, and related methods exploit future visual prediction to encourage temporal scene modeling (Li et al. 2025a; Zhang et al. 2025; Xia et al. 2026; Li et al. 2026c). OneVL similarly incorporates future visual information into compact internal representations to support downstream planning (Lu et al. 2026). Although these approaches provide richer temporal supervision, they mainly operate in the perspective-view image space and therefore emphasize appearance changes in 2D observations rather than explicitly constraining the underlying 3D spatial structure. WoTE and X-Mind instead model future scene evolution in the Bird's-Eye-View (BEV) space (Li et al. 2025b; Zhao et al. 2026), providing more structured scene-level information but inevitably compressing the physical world onto the ground plane.

Occupancy-based methods provide a more explicit representation of 3D scene geometry and temporal evolution. OccWorld, OccSora, DOME, I2-World, GenieDrive, OccLLaMA, and Occ-LLM model scene dynamics through occupancy prediction or forecasting (Zheng et al. 2024; Wang et al. 2026b; Gu et al. 2024; Liao et al. 2025b; Yang et al. 2026b; Wei et al. 2024; Xu et al. 2025). However, many of these methods treat occupancy as an input or prediction target of a dedicated world-modeling pipeline rather than as supervision for shaping the internal representations of a pretrained VLM. OccVLA takes multi-view images as input and aligns VLM representations with current occupancy (Liu et al. 2025), providing explicit supervision for 3D scene geometry. Nevertheless, it focuses on the current scene and does not explicitly model future scene evolution, which is critical for capturing physical dynamics and supporting downstream planning. PhysDrive instead

uses both current and future occupancy as structured training-time supervision, grounding VLM representations in both 3D geometry and dynamic scene evolution while retaining image and language inputs for downstream driving understanding and planning.

3 METHOD

The overall framework of **PhysDrive** is shown in Figure 2. Given multi-view images and textual instructions, PhysDrive generates natural-language responses for driving-understanding tasks or structured waypoints for motion planning. The key idea of PhysDrive is to physically ground the internal representations of the vision-language model (VLM) with structured 4D scene supervision. We define 4D physical grounding as jointly constraining the model representations with the current 3D geometry and the future temporal evolution of the driving scene. To this end, PhysDrive introduces dedicated physical grounding tokens whose hidden states interact with the visual representations and are jointly shaped by current and future occupancy supervision. For driving-understanding tasks, we further introduce optional region-of-interest (RoI) supervision to encourage question-relevant region grounding. We additionally introduce reinforcement learning to align generated trajectories with driving-oriented objectives.

3.1 4D PHYSICALLY GROUNDED REPRESENTATION LEARNING

Given multi-view images $\mathcal{I} = \{I_1, I_2, \ldots, I_N\}$ and a language instruction q, PhysDrive inserts a fixed number of special tokens between the multimodal input and the task output. The token sequence is enclosed by `<|latent_start|>` and `<|latent_end|>`, with a fixed number of `<|latent_pad|>` tokens. We refer to these tokens as *physical grounding tokens*, whose role is to encode 4D scene information and subsequently support downstream driving tasks. Physical grounding in driving requires capturing both the spatial structure of the current environment and how the scene evolves over time. We therefore divide the physical grounding tokens into two groups:

$$\mathbf{T}_{\text{phys}} = [\mathbf{T}_{\text{geo}}, \mathbf{T}_{\text{dyn}}],$$

where $\mathbf{T}_{\text{geo}}$ encodes 3D geometric information and $\mathbf{T}_{\text{dyn}}$ encodes dynamic scene information. For driving-understanding tasks, we additionally introduce a group of RoI tokens $\mathbf{T}_{\text{RoI}}$ to capture task-relevant spatial information. We extract the last-layer hidden states corresponding to the visual and physical grounding tokens:

$$[\mathbf{Z}_{\text{img}}, \mathbf{Z}_{\text{geo}}, \mathbf{Z}_{\text{dyn}}] = \text{VLM}_{\text{last}}(\mathcal{I}, q, \mathbf{T}_{\text{phys}}). \tag{1}$$

Here, $\mathbf{Z}_{\text{img}}$ denotes the visual hidden states, while $\mathbf{Z}_{\text{geo}}$ and $\mathbf{Z}_{\text{dyn}}$ denote the hidden states associated with the geometric and dynamic grounding tokens, respectively. These hidden states will interact with the visual features and be explicitly grounded by 4D occupancy supervision.

Spatial Grounding via Current Occupancy. Accurate driving understanding and planning require 3D spatial awareness; we therefore use current semantic occupancy to ground $\mathbf{Z}_{\text{geo}}$ in scene geometry. Specifically, we fuse the VLM visual hidden states $\mathbf{Z}_{\text{img}}$ with $\mathbf{Z}_{\text{geo}}$ using an attention operation and apply a depth prediction head D_{depth} on the fused features:

$$\mathbf{P}_d = \text{Softmax}(D_{\text{depth}}(\text{Attn}(\mathbf{Z}_{\text{img}}, \mathbf{Z}_{\text{geo}}))), \tag{2}$$

where $\mathbf{P}_d \in \mathrm{R}^{B\times N\times D\times H\times W}$ denotes the depth-bin probability volume. It is supervised by sparse LiDAR depth using a cross-entropy (CE) loss, denoted as $\mathcal{L}_{\text{depth}}$. Based on the predicted depth distribution, the visual features are lifted into an ego-centric 3D feature volume $\mathbf{F}_{\text{occ}}$. We then decode $\mathbf{F}_{\text{occ}}$ with D_{occ} into semantic occupancy: $\hat{\mathbf{O}} = D_{\text{occ}}(\mathbf{F}_{\text{occ}})$, where $\hat{\mathbf{O}} \in \mathrm{R}^{B\times C_{\text{occ}}\times Z\times H\times W}$ denotes the predicted occupancy volume. It is supervised using the ground-truth current occupancy $\mathbf{O}$:

$$\mathcal{L}_{\text{geo}} = \text{CE}\left(\hat{\mathbf{O}}, \mathbf{O}\right) + \lambda_{\text{lov}}\mathcal{L}_{\text{lov}}\left(\hat{\mathbf{O}}, \mathbf{O}\right) + \mathcal{L}_{\text{depth}}, \tag{3}$$

which combines CE loss and Lovasz-softmax loss (Berman et al. 2018). This supervision explicitly grounds the VLM hidden states in the 3D geometry of the observed scene.

Temporal Grounding via Future Occupancy. Spatial grounding alone describes only the current state of the environment. Autonomous driving additionally requires understanding how the scene evolves over time, since the motion and interaction of surrounding agents directly affect planning. We therefore use future occupancy to ground $\mathbf{Z}_{\mathrm{dyn}}$ in the dynamic evolution of the driving scene. Specifically, we fuse $\mathbf{Z}_{\mathrm{dyn}}$ with the current 3D feature volume using attention:

$$\mathbf{F}_{\mathrm{occ}}^{\mathrm{future}} = \mathrm{Attn}\left(\mathbf{F}_{\mathrm{occ}}, \mathbf{Z}_{\mathrm{dyn}}\right). \tag{4}$$

The resulting feature is decoded into future semantic occupancy using D_{occ}. The predicted future occupancy is supervised with the corresponding ground-truth occupancy using the same occupancy objective as above, denoted as $\mathcal{L}_{\mathrm{dyn}}$. Current and future occupancy together constitute our 4D physical grounding objective: $\mathcal{L}_{\mathrm{4D}} = \mathcal{L}_{\mathrm{geo}} + \mathcal{L}_{\mathrm{dyn}}$. The two terms provide complementary spatial and temporal supervision, jointly grounding the VLM representations in the 3D geometry and temporal evolution of the driving scene.

Region Grounding via RoI Supervision. While 4D physical grounding captures global scene geometry and dynamics, driving-understanding tasks often require the model to focus on regions that are particularly relevant to a given question. We therefore introduce RoI supervision to provide question-relevant region grounding. For understanding tasks, we activate a group of RoI tokens $\mathbf{T}_{\mathrm{RoI}}$ and obtain their last-layer hidden states $\mathbf{Z}_{\mathrm{RoI}}$. We fuse the visual hidden states with the RoI hidden states using attention: $\mathbf{Z}_{\mathrm{img}}^{\mathrm{RoI}} = \mathrm{Attn}\left(\mathbf{Z}_{\mathrm{img}}, \mathbf{Z}_{\mathrm{RoI}}\right)$, where the fused visual features are lifted into 3D space using the predicted depth distribution $\mathbf{P}_d$ to obtain the RoI feature $\mathbf{F}_{\mathrm{RoI}}$. We further decode $\mathbf{F}_{\mathrm{RoI}}$ into a binary region-relevance map using D_{RoI}, where 1 indicates question-relevant regions and 0 denotes the background. The prediction is supervised with the ground-truth question-relevant region mask $\mathbf{M}$:

$$\mathcal{L}_{\mathrm{RoI}} = \mathcal{L}_{\mathrm{BCE}}\left(D_{\mathrm{RoI}}\left(\mathbf{F}_{\mathrm{RoI}}\right), \mathbf{M}\right). \tag{5}$$

Grounded Representations for Understanding and Planning. The proposed supervision shapes the internal VLM representations from two complementary perspectives. The 4D physical grounding objective captures the spatial geometry and temporal dynamics of the driving scene, while RoI supervision provides region grounding for understanding tasks. PhysDrive retains a unified autoregressive generation framework, where the model produces natural-language responses for driving-understanding tasks and structured waypoints for planning tasks. The model is jointly optimized with the following objective:

$$\mathcal{L} = \mathcal{L}_{\mathrm{vlm}} + \mathcal{L}_{\mathrm{ground}}, \tag{6}$$

where $\mathcal{L}_{\mathrm{vlm}}$ denotes the standard next-token prediction loss and $\mathcal{L}_{\mathrm{ground}} = \mathcal{L}_{\mathrm{4D}} + \mathcal{L}_{\mathrm{RoI}}$. Through this joint optimization, the hidden states used for downstream understanding and planning tasks are explicitly grounded in the 4D driving scene.

3.2 Reinforced Planning Alignment

During supervised fine-tuning, PhysDrive learns physically grounded representations through 4D occupancy supervision while fitting expert trajectories with next-token prediction. To further align planning outputs with driving-oriented objectives beyond imitation learning, we adopt group-relative policy optimization (GRPO) (Guo et al. 2025a). For each prompt, PhysDrive samples multiple waypoint trajectories and evaluates them using $R(y) = R_{\mathrm{fmt}} + R_{\mathrm{l2}} + R_{\mathrm{env}}$, where R_{fmt} rewards valid waypoint formatting and $R_{\mathrm{l2}} = \exp(-\mathrm{ADE}/\sigma)$ encourages consistency with expert trajectories. The environmental reward R_{env} further accounts for driving quality. For nuScenes, we use collision avoidance and drivable-area compliance, i.e., $R_{\mathrm{env}} = R_{\mathrm{col}} + R_{\mathrm{dac}}$. For NAVSIM, we directly adopt the official PDMS (Dauner et al. 2024), which jointly evaluates no-at-fault collision, drivable-area compliance, time-to-collision, ego progress, and comfort. Importantly, physical grounding is jointly maintained during policy optimization. We therefore retain the grounding objective together with the GRPO objective and KL regularization:

$$\mathcal{L}_{\mathrm{RL}} = \mathcal{L}_{\mathrm{GRPO}} + \mathcal{L}_{\mathrm{KL}} + \mathcal{L}_{\mathrm{ground}}. \tag{7}$$

This allows PhysDrive to improve planning alignment while preserving the physically grounded representations learned during supervised training.

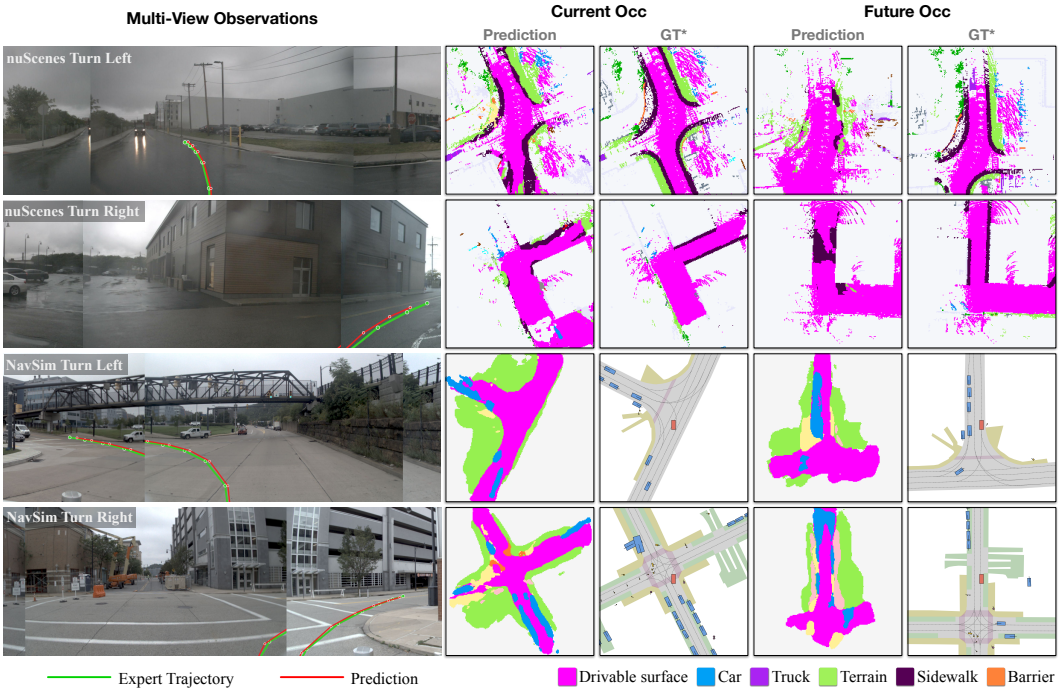

Figure 3: **Visualization of 4D physically grounded representations and planning.** We decode the grounded representations into current and future occupancy and visualize the corresponding planning trajectories across different benchmarks and navigation commands.

Table 1: Comparison on nuScenes open-loop planning benchmark. † denotes reproduced results.

Method	Ego History	L2 (m)↓				Collision (%)↓				Intersection (%)↓			
		1s	2s	3s	Avg.	1s	2s	3s	Avg.	1s	2s	3s	Avg.
Traditional Modular Paradigm													
UniAD (Hu et al. 2023)	√	0.20	0.42	0.75	0.46	0.02	0.25	0.84	0.37	0.20	1.33	3.24	1.59
VAD-Base (Jiang et al. 2023)	√	0.17	0.34	0.60	0.37	0.04	0.27	0.67	0.33	0.21	2.13	5.06	2.47
AD-MLP (Zhai et al. 2023)	√	0.15	0.32	0.59	0.35	0.00	0.27	0.85	0.37	0.27	2.52	6.60	2.93
UAD (Guo et al. 2025b)	√	0.13	0.28	0.48	0.30	0.00	0.19	0.61	0.27	0.13	1.08	2.89	1.37
World-Model based Paradigm													
Drive-WM (Wang et al. 2024)	√	0.43	0.77	1.20	0.80	0.10	0.21	0.48	0.26	–	–	–	–
Epona (Zhang et al. 2025)	√	0.61	1.17	1.98	1.25	0.01	0.22	0.85	0.36	–	–	–	–
DriveLaW (Xia et al. 2026)	√	0.44	1.10	1.91	1.15	0.15	**0.10**	0.48	0.24	–	–	–	–
VLM based Paradigm													
EMMA (Hwang et al. 2024)	–	**0.14**	0.29	0.54	0.32	–	–	–	–	–	–	–	–
RDA-Driver (Huang et al. 2024)	√	0.23	0.73	1.54	0.80	**0.00**	0.13	0.83	0.32	–	–	–	–
DriveVLM (Tian et al. 2024)	√	0.18	0.34	0.68	0.40	0.10	0.22	**0.45**	0.27	–	–	–	–
ORION (Fu et al. 2025)	√	0.17	0.31	0.55	0.34	0.05	0.25	0.80	0.37	–	–	–	–
DrivePI (Liu et al. 2026)	–	0.24	0.46	0.78	0.49	0.38	0.27	0.48	0.38				
OmniDrive-Q++ (Wang et al. 2025)	√	**0.14**	0.29	0.55	0.33	**0.00**	0.13	0.78	0.30	0.56	2.48	5.96	3.00
VGGDrive† (Wang et al. 2026a)	–	0.15	0.32	0.60	0.36	0.02	0.18	0.90	0.37	0.37	1.88	4.94	2.40
SpaceDrive+ (Li et al. 2026b)	√	0.15	0.29	**0.51**	0.32	0.04	0.18	0.49	0.23	**0.22**	**0.80**	2.79	1.27
PhysLaD-3B	–	**0.14**	**0.28**	**0.51**	**0.31**	0.04	**0.12**	0.51	0.22	0.31	1.15	2.79	1.42
PhysLaD-3B + RL	–	**0.14**	**0.28**	**0.51**	**0.31**	**0.00**	**0.12**	0.47	**0.20**	0.23	0.94	**2.25**	**1.14**

4 EXPERIMENTS

Experimental Setup. We use Qwen2.5-VL-3B (Bai et al. 2025) as the base models. Driving understanding is evaluated on NuInstruct (Ding et al. 2024), while planning is evaluated on nuScenes and NAVSIM (Dauner et al. 2024). We use all six camera views for NuInstruct and nuScenes, and

Table 2: Comparison on the NAVSIM navtest. The best result in each group is highlighted in **bold**, while the second-best result is underlined.

Method	NC↑	DAC↑	EP↑	TTC↑	Comf.↑	PDMS↑
TransFuser (Chitta et al. 2022)	97.78	92.63	78.88	92.89	99.98	83.88
PARA-Drive (Weng et al. 2024)	97.90	92.40	79.30	93.00	99.8	84.00
Hydra-MDP-V_{8192} (Li et al. 2024)	**98.30**	96.00	78.70	94.60	**100**	86.50
DRAMA (Yuan et al. 2024)	98.19	95.18	81.33	94.17	**100**	86.87
DiffusionDrive (Liao et al. 2025a)	98.20	**96.20**	**82.20**	**94.70**	**100**	**88.10**
World-Model based Paradigm						
Epona (Zhang et al. 2025)	97.90	95.10	80.40	93.80	99.9	86.20
ReSim (Yang et al. 2026a)	–	–	–	–	–	86.60
DriveVLA-W0 (Li et al. 2026c)	**98.40**	**95.30**	**80.90**	**95.20**	**100**	**87.20**
VLM based Paradigm						
AutoVLA (Text Waypoint)	–	–	–	–	–	71.31
ReCogDrive (Text Waypoint)	–	–	–	–	–	84.10
FSDrive (Zeng et al. 2026)	98.20	93.80	80.10	93.30	99.90	85.10
AdaThinkDrive (SFT) (Luo et al. 2025)	98.50	94.40	79.90	94.90	**100**	86.20
ImagiDrive (Li et al. 2025a)	97.90	95.50	80.70	93.10	99.9	86.40
PhysLaD-3B	98.70	95.22	81.09	95.86	99.99	87.52
PhysLaD-7B	98.58	95.52	81.55	95.80	99.98	87.85
PhysLaD-3B + RL	**98.96**	**97.65**	**83.27**	**96.90**	99.98	**90.25**



Figure 4: **Effect of reinforcement learning on planning alignment.** RL improves SFT trajectories toward the expert trajectories; the upper-left values denote the PDMS change before → after RL.

`B0`, `F0`, `L0`, `L1`, `R0`, and `R1` for NAVSIM. Occupancy annotations are obtained from Occ3D (Tian et al. 2023) for nuScenes and UniSceneV2 (Li et al. 2026a) for NAVSIM. Each representation-token group contains 32 tokens by default. During SFT, we train the VLM and auxiliary decoders for four epochs using a cosine schedule, with learning rates of 2×10^{-5} and 2×10^{-4}, respectively. All experiments are conducted on 16 GPUs with 80 GB of VRAM each.

Physically Grounded Representation Visualization. Figure 3 visualizes the grounded representations by decoding them into current and future occupancy, together with the corresponding predicted trajectories. The decoded current and future occupancy qualitatively recover the spatial structure and dynamic evolution of the driving scene, while the predicted trajectories remain close to the expert trajectories. These results provide qualitative evidence that the physically grounded representations encode information about both current 3D geometry and future scene dynamics, supporting accurate and physically compliant planning.

nuScenes Open-Loop Planning. Our method takes multi-view images as visual input and encodes the navigation command in the textual instruction, without using historical ego waypoints. Planning



Figure 5: **Qualitative comparison on the NuInstruct understanding benchmark.** We compare our method with VGGDrive and Qwen2.5-VL across different understanding tasks.

Table 3: Comparison on NuInstruct understanding benchmark. † denotes reproduced results.

Method	*Perception*						*Prediction*		*Risk*	*Planning*
	Dis↓	Sped↓	Ins↓	Clos↑	Sta↑	SameR↑	Mot↓	Sta↑	Avg.↑	BLEU↑
MV-MLLM	28.9	6.2	2.4	27.9	19.6	30.9	9.3	44.3	-	31.4
BEV-InMLLM	24.5	3.5	4.2	31.6	19.0	34.6	4.1	44.7	-	35.2
VGGDrive†	5.5	2.8	1.6	37.0	50.7	48.6	3.6	55.2	14.4	59.2
Qwen-2.5VL-3B	4.8	3.1	1.7	43.3	51.6	51.9	3.6	57.0	24.8	61.4
Ours-3B	**2.5**	**2.8**	**1.6**	**43.5**	**71.9**	**52.9**	**1.8**	**75.0**	**33.2**	**67.7**
Δ	-2.3	-0.3	-0.1	+0.2	+20.3	+1.0	-1.8	+18.0	+8.4	+6.3
Qwen-2.5VL-7B	3.9	2.6	**1.6**	43.8	34.7	53.7	3.3	66.8	31.1	66.5
Ours-7B	**2.4**	**2.4**	**1.6**	**45.9**	**76.2**	**53.8**	**1.7**	**77.4**	**32.4**	**69.2**
Δ	-1.5	-0.2	0.0	+2.1	+41.5	+0.1	-1.6	+10.6	+1.3	+2.7

Table 4: Ablation studies on grounding supervision and the number of grounding tokens.

Setting	Understanding				Planning		
	PP.↑	MAE↓	Risk↑	BLEU↑	L2↓	Coll.↓	Inter.↓
Baseline	48.7	3.9	24.8	61.4	0.33	0.30	2.25
+ Cur. Occ.	50.9	3.1	31.8	67.7	0.33	0.29	1.85
+ Cur. Occ. + ROI	64.4	3.0	32.2	67.8	–	–	–
+ Cur. Occ. + Fut. Occ.	63.8	3.0	31.3	**67.9**	**0.31**	**0.22**	**1.42**
Full (32 latent tokens)	**65.5**	**2.1**	**33.2**	67.7	**0.31**	**0.22**	**1.42**
Full (8 latent tokens)	63.5	3.1	31.2	67.5	0.32	0.24	2.23
Full (16 latent tokens)	64.4	3.0	**33.8**	**68.5**	0.35	0.23	2.05
Full (64 latent tokens)	**65.2**	**2.8**	33.2	68.1	**0.31**	**0.21**	**1.70**

performance is evaluated using the L_2 displacement error, collision rate, and intersection rate at 1, 2, and 3 seconds, as reported in Table 1. After Stage 1 physically grounded supervised fine-tuning, PhysDrive achieves the lowest L_2 error and collision rate, while maintaining a competitive intersection rate compared with SpaceDrive+, which additionally utilizes ego-history information. These results indicate that physically grounded VLM representations help encode the 3D structure of the driving scene and its possible future evolution, leading to more accurate trajectory prediction, fewer collisions, and fewer departures from the drivable area. Building on Stage 1, our proposed RL framework further reduces the average collision rate from 0.22% to 0.20% and the intersection rate from 1.42% to 1.14%, while maintaining an average L_2 error of 0.31 m, achieving the best overall performance across all three metrics.

NAVSIM Planning. We evaluate PhysDrive on the NAVSIM navtest using no-at-fault collision (NC), drivable-area compliance (DAC), ego progress (EP), time-to-collision (TTC), comfort, and predictive driver model score (PDMS), as reported in Table 2. Compared with VLM-based methods that also generate textual waypoints, PhysDrive-3B substantially outperforms AutoVLA and RecogDrive by 16.21 and 3.42 PDMS points, respectively. PhysDrive also achieves clear gains over methods equipped with other forms of auxiliary supervision, including AdaThinkDrive with textual CoT supervision, FSDrive with future-image supervision, and ImagiDrive with an auxiliary world model. These results suggest that physically grounding the VLM representations with 4D scene supervision provides an effective basis for downstream planning. Within PhysDrive, scaling the base model from 3B to 7B improves PDMS from 87.52 to 87.85, while reinforcement learning further raises the 3B model from 87.52 to 90.25 through substantial improvements in NC, DAC, EP, and TTC. Figure 4 presents several representative examples that achieve substantial PDMS improvements after RL training. In the first scene, the SFT trajectory passes too close to vehicles along the road, resulting in a collision and a PDMS of zero. After RL, the predicted trajectory keeps a safer distance from surrounding vehicles and successfully reaches the target position, achieving a full PDMS score. The remaining examples further illustrate how the policy learns to slow down appropriately while waiting and to remain within the drivable area. These results suggest that 4D physical grounding provides informative representations for downstream driving tasks, while RL further translates these representations into safer and more driving-compliant planning behaviors.

Driving Scenario Understanding. We evaluate driving scenario understanding on NuInstruct, which includes discriminative, numerical prediction, planning, and risk-grounding tasks. We use accuracy for discriminative tasks, mean absolute error (MAE) for numerical predictions, and BLEU for planning tasks that require predicting the driving decision. Since the official risk-grounding evaluation code is unavailable, we adopt a simplified accuracy metric: a prediction is considered correct only if it identifies the correct camera view and object class and achieves an intersection over union (IoU) greater than 0.5 with the ground-truth bounding box. As shown in Table 3, PhysDrive outperforms Qwen2.5-VL, VGGDrive, MV-MLLM, and BEV-InMLLM, with notable gains in perception distance and status, prediction status, risk grounding, and planning. Figure 5 further shows that PhysDrive correctly grounds the closest vehicle and the vehicle traveling on the same road, produces a more accurate distance estimate, and identifies the pedestrian posing a potential risk, whereas the compared models make localization or estimation errors or overlook the hazard. These results demonstrate that physically grounding the internal VLM representations benefits not only planning but also diverse driving-understanding tasks.

Ablation on Grounding Supervision. We ablate the contribution of different grounding signals to driving understanding and planning, as reported in Table 4. For understanding, we report Perception and Prediction accuracy, abbreviated as PP., for the discriminative perception and prediction tasks. Current occupancy supervision notably reduces MAE from 3.9 to 3.1, improves risk-grounding accuracy from 24.8 to 31.8, and raises planning BLEU from 61.4 to 67.7, demonstrating the benefit of grounding VLM representations in the current 3D scene geometry. Building on current occupancy supervision, RoI supervision substantially improves PP accuracy from 50.9 to 64.4 by encouraging question-relevant region grounding. Future occupancy supervision achieves the highest BLEU of 67.9 among the individual variants, indicating that modeling dynamic scene evolution benefits downstream driving-decision prediction. Combining all grounding signals yields the strongest overall understanding performance, including 65.5 PP accuracy, 2.1 MAE, and 33.2 risk-grounding accuracy. For planning, RoI supervision is not applied, as planning relies more on global scene evolution than on localized question-relevant regions. Current occupancy supervision reduces the intersection rate from 2.25 to 1.85, while introducing future occupancy supervision further lowers it to 1.42 and reduces the collision rate from 0.30 to 0.22, with little change in the L_2 error. These results demonstrate the complementary roles of current and future occupancy: current occupancy strengthens the representation of 3D scene geometry, while future occupancy provides dynamic information that is particularly important for collision avoidance and drivable-area compliance.

Ablation on the Number of Grounding Tokens. We further investigate the effect of the number of grounding tokens on driving understanding and planning. Increasing the number of grounding tokens generally improves overall performance. Compared with 8 and 16 tokens, using 32 tokens achieves the best PP accuracy and MAE while substantially reducing collision and intersection rates. Although 16 tokens yield slightly higher risk-grounding accuracy, the 32-token setting provides stronger overall performance across tasks. Further increasing the number to 64 brings only marginal or inconsistent gains: BLEU and collision rate improve slightly, whereas PP accuracy, MAE, and intersection rate degrade. We therefore adopt 32 grounding tokens as the default setting, offering a favorable trade-off between downstream performance and computational cost.

5 Conclusion

We presented **PhysDrive**, a physically grounded vision-language model for autonomous driving. PhysDrive leverages 4D occupancy supervision to shape VLM representations with both current 3D scene geometry and future dynamic evolution, complemented by optional RoI supervision for question-relevant regions. Through this grounding, the model learns to internalize 4D physical cues from raw image observations, providing more informative representations for downstream driving understanding and planning. A GRPO-based reinforcement learning stage further aligns planning outputs with driving-oriented objectives while preserving the learned physical grounding. Extensive experiments on NuInstruct, nuScenes, and NAVSIM demonstrate consistent improvements in both driving understanding and planning. Overall, PhysDrive highlights the importance of grounding visual-language representations in structured 4D scene information, offering a practical way to better connect visual perception with downstream driving decisions.

AI USE DISCLOSURE

In this work, we did not use generative AI tools for any tasks requiring mandatory disclosure, including research methodology or experiment design, method implementation, data processing, theoretical development, or result interpretation. We used generative AI tools solely to edit the manuscript for English grammar, clarity, and readability. All AI-assisted edits were carefully reviewed by the authors, who take full responsibility for the final content of this work.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE transactions on pattern analysis and machine intelligence*, 45(11):12878–12895, 2022.

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. *Advances in Neural Information Processing Systems*, 2024.

Xinpeng Ding, Jianhua Han, Hang Xu, Xiaodan Liang, Wei Zhang, and Xiaomeng Li. Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Songen Gu, Wei Yin, Bu Jin, Xiaoyang Guo, Junming Wang, Haodong Li, Qian Zhang, and Xiaoxiao Long. Dome: Taming diffusion model into high-fidelity controllable occupancy world model. *arXiv preprint arXiv:2410.10429*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025a.

Mingzhe Guo, Zhipeng Zhang, Yuan He, Ke Wang, Liping Jing, and Haibin Ling. End-to-end autonomous driving without costly modularization and 3d manual annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b.

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*, 2023.

Zhijian Huang, Tao Tang, Shaoxiang Chen, Sihao Lin, Zequn Jie, Lin Ma, Guangrun Wang, and Xiaodan Liang. Making large language models better planners with reasoning-decision alignment. In *European Conference on Computer Vision*. Springer, 2024.

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.

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

Bohan Li, Xin Jin, Hu Zhu, Hongsi Liu, Ruikai Li, Jiazhe Guo, Kaiwen Cai, Chao Ma, Yueming Jin, Hao Zhao, et al. Scaling up occupancy-centric driving scene generation: Dataset and method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026a.

Jingyu Li, Bozhou Zhang, Xin Jin, Jiankang Deng, Xiatian Zhu, and Li Zhang. Imagidrive: A unified imagination-and-planning framework for autonomous driving. *arXiv preprint arXiv:2508.11428*, 2025a.

Peizheng Li, Zhenghao Zhang, David Holtz, Hang Yu, Yutong Yang, Yuzhi Lai, Rui Song, Andreas Geiger, and Andreas Zell. Spacedrive: Infusing spatial awareness into vlm-based autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2026b.

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*, 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. In *International Conference on Learning Representations*, 2026c. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/0d70423f59c5fdd24f0dd3fa52e34623-Paper-Conference.pdf.

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. In *International Conference on Learning Representations*, 2026d. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/ff7bf6014f7826da531aa50f4538ee19-Paper-Conference.pdf.

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, and Xinggang Wang. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, June 2025a.

Zhimin Liao, Ping Wei, Ruijie Zhang, Shuaijia Chen, Haoxuan Wang, and Ziyang Ren. I2-world: Intra-inter tokenization for efficient dynamic 4d scene forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025b.

Ruixun Liu, Lingyu Kong, Derun Li, and Hang Zhao. Occvla: Vision-language-action model with implicit 3d occupancy supervision. *arXiv preprint arXiv:2509.05578*, 2025.

Zhe Liu, Runhui Huang, Rui Yang, Siming Yan, Zining Wang, Lu Hou, Di Lin, Xiang Bai, and Hengshuang Zhao. Drivepi: Spatial-aware 4d mllm for unified autonomous driving understanding, perception, prediction and planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2026.

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*, 2026.

Yuechen Luo, Fang Li, Shaoqing Xu, Zhiyi Lai, Lei Yang, Qimao Chen, Ziang Luo, Zixun Xie, Shengyin Jiang, Jiaxin Liu, et al. Adathinkdrive: Adaptive thinking via reinforcement learning for autonomous driving. *arXiv preprint arXiv:2509.13769*, 2025.

Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in Neural Information Processing Systems*, 2023.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024.

Jie Wang, Guang Li, Zhijian Huang, Chenxu Dang, Hangjun Ye, Yahong Han, and Long Chen. Vg-gdrive: Empowering vision-language models with cross-view geometric grounding for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2026a.

Lening Wang, Wenzhao Zheng, Yilong Ren, Han Jiang, Zhiyong Cui, Haiyang Yu, and Jiwen Lu. Occsora: 4d occupancy generation models as world simulators for autonomous driving. *IEEE Transactions on Image Processing*, 2026b.

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 *Proceedings of the computer vision and pattern recognition conference*, 2025.

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

Julong Wei, Shanshuai Yuan, Pengfei Li, Qingda Hu, Zhongxue Gan, and Wenchao Ding. Occllama: An occupancy-language-action generative world model for autonomous driving. *arXiv preprint arXiv:2409.03272*, 2024.

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*, 2024.

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

Tianshuo Xu, Hao Lu, Xu Yan, Yingjie Cai, Bingbing Liu, and Yingcong Chen. Occ-llm: Enhancing autonomous driving with occupancy-based large language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025.

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*, 2026a.

Zhenya Yang, Zhe Liu, Yuxiang Lu, Liping Hou, Chenxuan Miao, Siyi Peng, Bailan Feng, Xiang Bai, and Hengshuang Zhao. Geniedrive: Towards physics-aware driving world model with 4d occupancy guided video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2026b.

Chengran Yuan, Zhanqi Zhang, Jiawei Sun, Shuo Sun, Zefan Huang, Christina Dao Wen Lee, Dongen Li, Yuhang Han, Anthony Wong, Keng Peng Tee, et al. Drama: An efficient end-to-end motion planner for autonomous driving with mamba. *arXiv preprint arXiv:2408.03601*, 2024.

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*, 2026.

Jiang-Tian Zhai, Ze Feng, Jinhao Du, Yongqiang Mao, Jiang-Jiang Liu, Zichang Tan, Yifu Zhang, Xiaoqing Ye, and Jingdong Wang. Rethinking the open-loop evaluation of end-to-end autonomous driving in nuscenes, 2023.

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*, 2025.

Bohao Zhao, Chengrui Wei, Guangfeng Jiang, Ruixin Liu, Xuejie Lv, Liu Liang, Sutao Deng, Xiuyang Fan, Pengkun Zheng, Jinyun Zhou, et al. X-mind: Efficient visual chain-of-thought via predictive world model for end-to-end driving. *arXiv preprint arXiv:2606.28758*, 2026.

Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. In *European conference on computer vision*. Springer, 2024.

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

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

A APPENDIX

A.1 ADDITIONAL REINFORCEMENT LEARNING DETAILS

This section provides further details of the reinforcement learning stage of PhysDrive, including its optimization configuration, challenging-case data selection strategy, and qualitative effects on planning.

Training Setting. During reinforcement learning, we freeze the vision encoder, auxiliary decoders, and LLM token embeddings, while optimizing only the Transformer layers of the LLM. We train for 10 epochs using a constant learning rate of 1×10^{-6}. For each prompt, we sample four trajectories, and the per-device training batch size is set to 8. The Kullback–Leibler (KL) coefficient β is set to 0.002, and the sampling temperature is set to 1.5. We implement group-relative policy optimization (GRPO) in-house, use vLLM to accelerate rollout generation, and adopt DeepSpeed ZeRO-2 for distributed optimization. Updated policy weights are transferred to the vLLM inference engine using CUDA inter-process communication (IPC).

Data Selection. The NAVSIM training split contains 101,602 samples. We only perform RL on challenging scenarios for efficient training. Specifically, we first perform supervised fine-tuning on the full training split for four epochs and evaluate the resulting policy on this split solely for sample selection. We then retain samples with a per-sample PDMS below 0.8. These cases typically involve collisions, departures from the drivable area, or insufficient ego progress (EP). This procedure yields 5,702 samples, corresponding to 5.61% of the full training set. The RL is executed for 10 epoch on these selected samples.

Effectiveness of Reinforcement Learning. Figure 6 presents representative cases with substantial PDMS gains after reinforcement learning. For DAC, RL corrects trajectories that drift toward lane boundaries or non-drivable regions; for NC, it steers the vehicle away from nearby agents to avoid collisions. For EP, RL extends prematurely terminated trajectories to make greater route progress, while for TTC, it adjusts the path around surrounding traffic to maintain safer clearance. These examples show that RL effectively corrects diverse SFT planning failures, improving drivable-area compliance, collision avoidance, progress, and interaction safety.

A.2 ADDITIONAL NUINSTRUCT VISUALIZATIONS

Figure 7 visualizes the physically grounded representations learned by PhysDrive through occupancy and region-of-interest (RoI) predictions. The decoded occupancy reflects the scene-level 3D geometry captured by the model, while the RoI predictions indicate the question-relevant regions emphasized for each query. Across all examples, the predicted RoIs remain compact and closely align with the annotated targets, showing that PhysDrive can accurately localize relevant objects rather than relying only on global scene context.

In the first nighttime example, PhysDrive correctly identifies the distant vehicle ahead as the closest object to the ego vehicle despite poor illumination and its small image footprint. In the second example, the predicted RoI accurately focuses on the front-right vehicle when queried about the closest object in front, while suppressing irrelevant vehicles and background regions. For the two same-road queries, PhysDrive focuses on vehicles within the ego vehicle's road corridor, indicating that the grounded representations capture both object information and its spatial relationship to the surrounding scene geometry. In the final approaching-object example, the predicted RoI concentrates on the nearby van on the right and closely matches the annotated target, demonstrating effective grounding of potentially interacting objects.

These visualizations further illustrate the complementary roles of occupancy and RoI supervision. Occupancy supervision grounds the VLM representations in global 3D scene geometry, while RoI supervision provides question-relevant region grounding. Together, they enable PhysDrive to combine scene-level spatial understanding with localized task-relevant evidence, supporting its performance across proximity, same-road, and approaching-object understanding tasks.

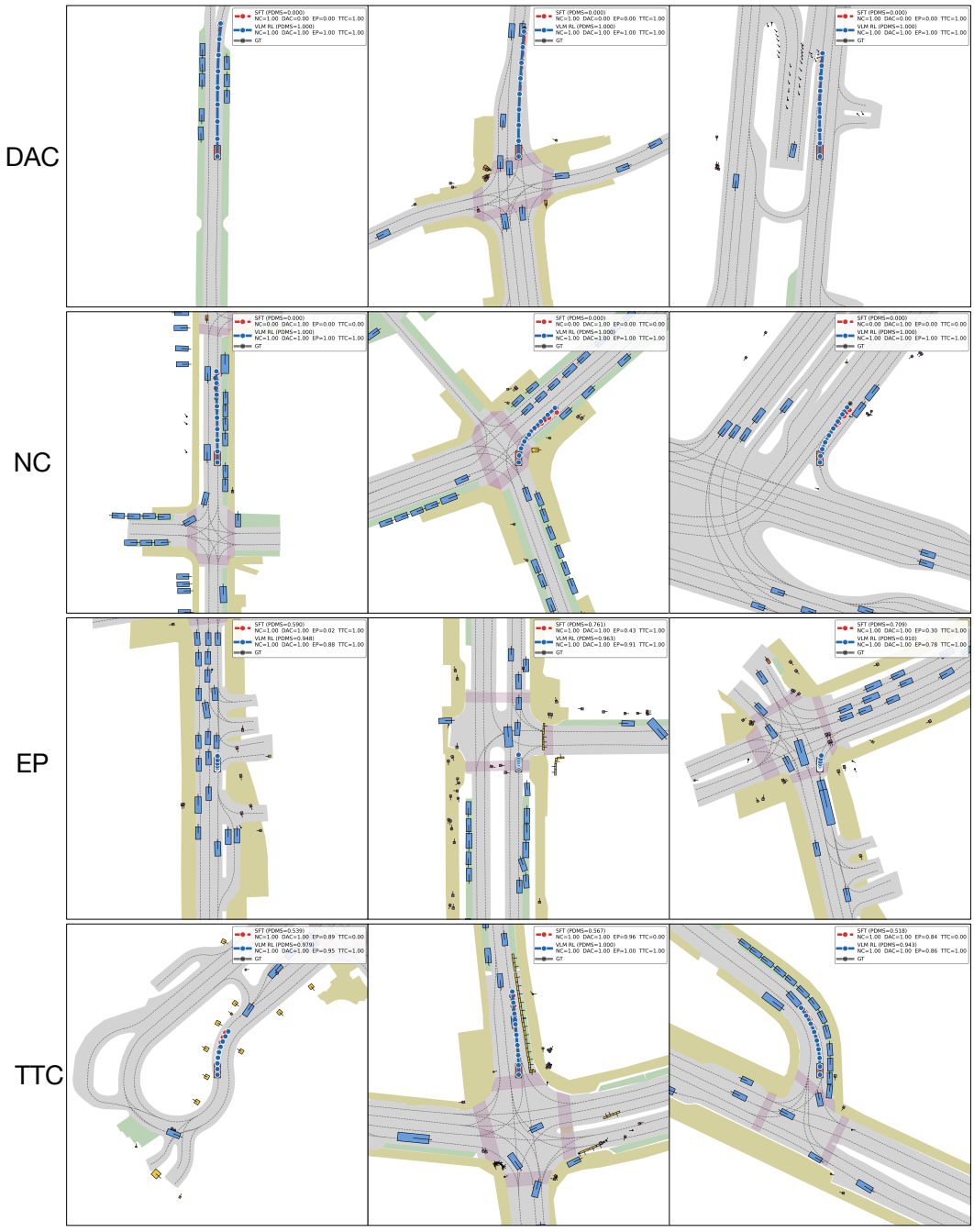

Figure 6: Representative cases with substantial PDMS improvements after reinforcement learning. Each row groups examples in which the PDMS gain is primarily driven by improvements in drivable-area compliance (DAC), no-at-fault collision (NC), ego progress (EP), or time-to-collision (TTC). Red, blue, and gray denote trajectories generated by the SFT policy, the RL-enhanced policy, and the ground truth, respectively.

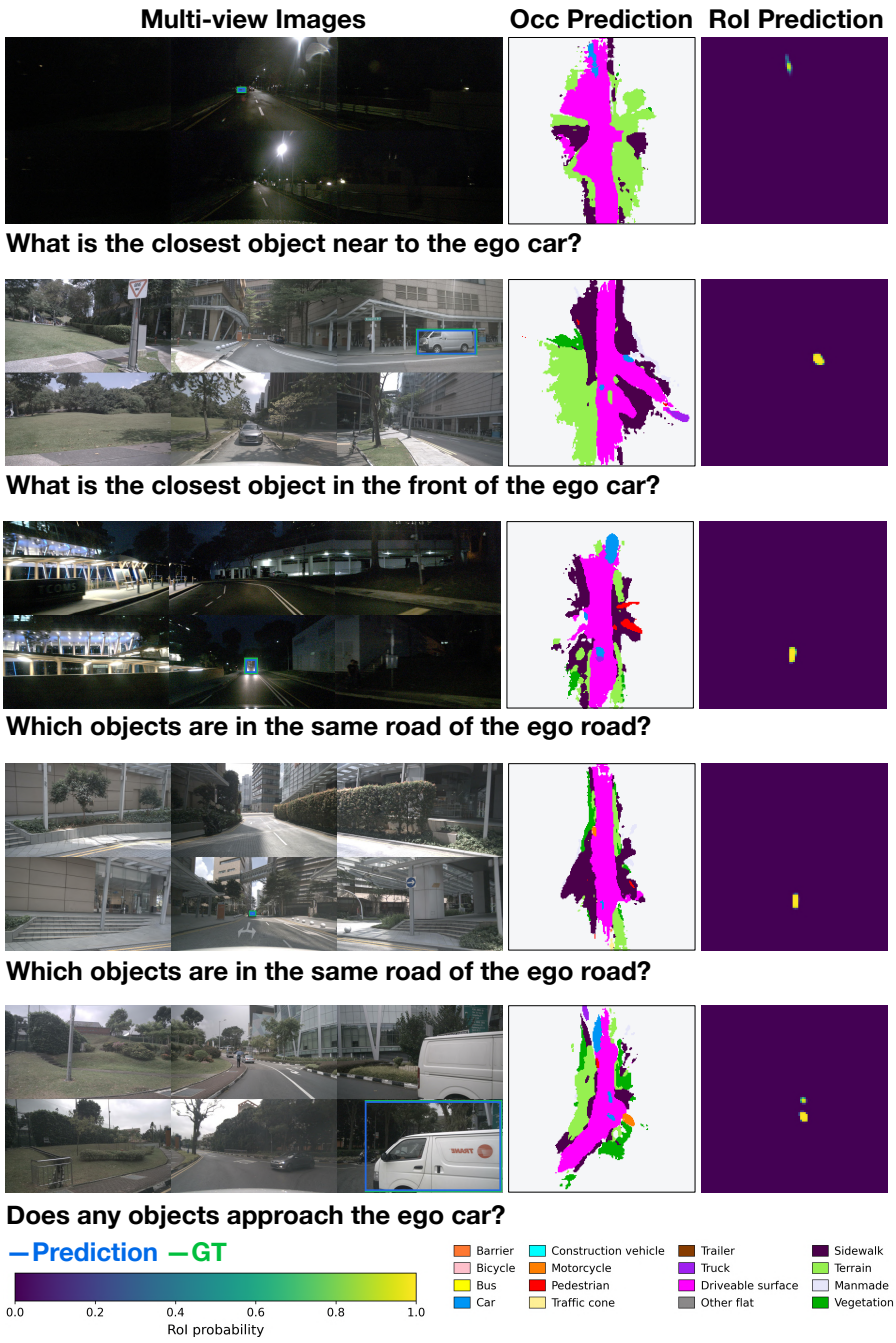

Figure 7: Visualization of physically grounded representations on NuInstruct. Decoded occupancy captures global 3D scene geometry, while query-conditioned RoI maps localize question-relevant objects. Their complementary grounding supports more accurate driving understanding. Blue and green boxes denote ground-truth and predicted objects, respectively.