DOE-3: WORLD-LANGUAGE-ACTION MODELS FOR AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

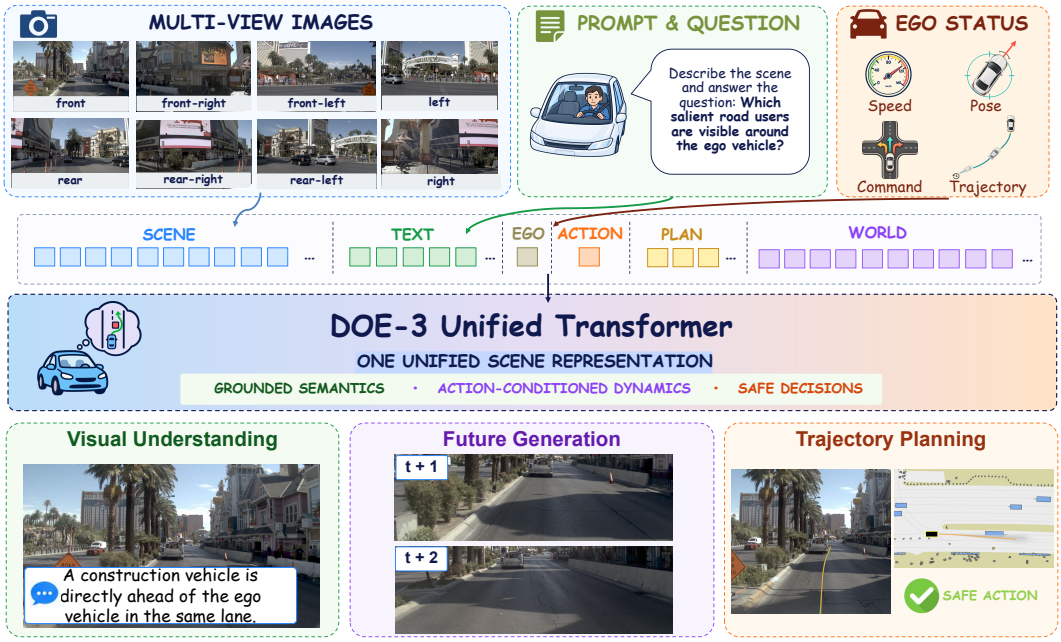


Figure 1: Doe-3: understand, predict, and act through a shared scene representation. Surround-view scene latents, language, and actions are jointly modeled in an autoregressive transformer, with language, planning, and future-prediction objectives jointly shaping the representation.

ABSTRACT

Autonomous driving requires understanding the observed scene, predicting the consequences of actions, and planning safe trajectories. Existing efforts to combine these capabilities often use separate visual representations for understanding and generation. We propose Doe-3, a world-language-action model that jointly models world states, language, and actions. We adopt the pretrained driving world representation to encode synchronized multi-view observations into scene tokens that retain appearance, semantics, and geometry. We introduce a shared autoregressive backbone that treats this world representation as a modality alongside language and action. The backbone provides a common temporal context for grounded language responses, trajectory planning, and action-conditioned future prediction. We attach a lightweight diffusion head that maps autoregressive world tokens to future scene latents, which pretrained decoders convert into RGB, depth, semantic, and occupancy outputs. We jointly train the backbone with language, planning, and future-prediction objectives to connect semantic understanding with scene dynamics. We adopt a three-stage schedule that progressively aligns the scene interface, adapts the backbone, and jointly optimizes the model. Doe-3 achieves a PDMS of 94.0 on NAVSIMv1 and an EPDMS of 90.8 on NAVSIMv2, outperforming the strongest prior methods. On nuScenes, it reduces the future-depth absolute relative error to 0.13 and reaches 27.03 mIoU and 61.49 IoU in camera-based 4D occupancy forecasting.

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.

1 INTRODUCTION

Autonomous driving has advanced rapidly with end-to-end models that optimize perception, prediction, and planning toward a shared driving objective (Hu et al., 2023b; Jiang et al., 2023). This paradigm provides a scalable path for learning from growing driving datasets, with data-scaling studies reporting performance gains and large multimodal models achieving strong planning results (Naumann et al., 2025; Hwang et al., 2024). Beyond scaling, effective planning also depends on whether the learned representation captures both the semantic meaning of the current scene and how this scene may evolve under different candidate ego actions.

Safe trajectory planning requires both understanding semantically relevant cues in the present scene and anticipating how the scene may evolve in response to ego actions (Hu et al., 2023b; Wang et al., 2024c). Vision-language-action (VLA) models emphasize language-grounded scene understanding for planning (Hwang et al., 2024; Zhou et al., 2025b), whereas world-action models (WAMs) emphasize action-conditioned future prediction for planning (Wang et al., 2024c; Xia et al., 2026). These two forms of grounding are complementary. Semantic understanding can resolve intent and instructions that are ambiguous from motion alone, while future prediction can anticipate interactions that are not explicit from the current observation. VLA models ground actions in visual and linguistic context but may miss interaction dynamics, while WAMs model scene evolution but may not resolve semantic intent, and our goal is to capture both in a shared scene representation. Recent efforts bridge these directions by either autoregressively generating image, language, and action tokens or combining autoregressive understanding with diffusion-based generation in a joint transformer (Zheng et al., 2024b; Zuo et al., 2026; NVIDIA, 2026). Other methods pass representations from an autoregressive backbone to a separate diffusion generator (Xiong et al., 2026; Yang et al., 2026). These designs mainly unify the computation, with discrete autoregressive models operating on 2D image codes, while diffusion-based designs couple semantic features to separate generative latents through shared attention or conditioning features (Zheng et al., 2024b; Zuo et al., 2026; NVIDIA, 2026; Xiong et al., 2026; Yang et al., 2026). This leaves a need for a unified driving model that captures the correlations among world states, language, and actions in one architecture.

To address this gap, we propose Doe-3, a world-language-action (WLA) model built on a pretrained driving world representation (Zhuo et al., 2026b), as shown in Figure 1. It maps synchronized multi-view images to scene tokens that retain appearance, semantics, and geometry. We introduce a shared autoregressive driving backbone that models the causal relationships among world, language, and action tokens. We attach language and planning readouts to this backbone and use its action-conditioned future-world tokens to model scene evolution. When we need dense future-scene outputs, we employ a lightweight diffusion refinement head on the world tokens to predict accurate world representations and use the pretrained decoders to recover RGB appearance, depth, semantic maps, and occupancy. We train future-scene prediction with latent-matching and flow-matching losses that supervise the autoregressive world tokens through the diffusion head. We optimize trajectory, proposal-dynamics, and score losses for planning and a language loss for scene descriptions and visual-question answering, which update the shared autoregressive backbone through its modality tokens. We first align the interface between the world representations and the backbone with the backbone and task heads frozen, then adapt the backbone with low-rank adaptation (LoRA) while keeping the heads frozen, and finally jointly tune the interface, LoRA, planner, and diffusion conditioning modules. Experiments on NAVSIM demonstrate the superiority of our model.

2 RELATED WORK

Vision-Language-Action Models for Driving. Vision-language-action models connect scene understanding and driving decisions through language supervision and pretrained multimodal knowledge (Wang et al., 2024a; Hwang et al., 2024; Jiang et al., 2024; Zhou et al., 2025b). These methods align visual features with language and train models to answer driving questions, reason about behavior, or predict trajectories (Wang et al., 2024a; Hwang et al., 2024; Zhou et al., 2025b). They support planning through either autoregressive action prediction or a trajectory predictor conditioned on language decisions (Hwang et al., 2024; Jiang et al., 2024; Zhou et al., 2025b). Their main emphasis is the mapping from observed scenes and semantic reasoning to driving actions. We extend this scope in Doe-3 by jointly training a shared autoregressive backbone for grounded language, planning, and action-conditioned future prediction. We use one driving world representation for observed scenes and future targets, linking semantic understanding to scene evolution.

World-Action Models for Driving. Driving world models learn scene dynamics from video and generate future observations under text, structural, or action conditions (Hu et al., 2023a; Wang et al., 2023; Wen et al., 2024). Autoregressive and diffusion architectures support controllable generation by modeling temporal evolution and preserving visual consistency across frames and camera views (Hu et al., 2023a; Wang et al., 2024c; Wen et al., 2024; Zhang et al., 2025). World-action models further connect future prediction to planning through imagined outcomes, shared video latents, or latent reasoning over action proposals (Wang et al., 2024c; Zhang et al., 2025; Xia et al., 2026; Tan et al., 2026). These approaches establish future prediction as a source of information for driving decisions. We build on this idea by using the pretrained DriveTok representation, which retains multi-view appearance, semantics, and geometry (Zhuo et al., 2026b). In Doe-3, a shared autoregressive backbone models world, language, and action tokens, while a lightweight diffusion head predicts future latents in the same representation space. This design supports grounded language and planning alongside future RGB, depth, semantic, and occupancy decoding.

Unified World-Language-Action Models. Recent methods combine understanding, generation, and action modeling through shared token sequences or coupled neural architectures (Zheng et al., 2024b; Chen et al., 2025; Zuo et al., 2026; NVIDIA, 2026; Xiong et al., 2026; Yang et al., 2026). Autoregressive approaches interleave discrete visual and action tokens, with language tokens providing an additional modality in language-capable models (Zheng et al., 2024b; Chen et al., 2025). Hybrid approaches combine autoregressive and diffusion components through shared attention or pass autoregressive features to diffusion generators (Zuo et al., 2026; NVIDIA, 2026; Xiong et al., 2026; Yang et al., 2026). Related efforts also share scene representations across 3D understanding and generation or across video prediction and planning (Zhou et al., 2025a; Xia et al., 2026). We bring these directions together by treating the pretrained DriveTok world representation as a modality alongside language and action in a shared autoregressive model. Our contribution centers on modeling the relationships among these modalities with a shared backbone and a lightweight diffusion refinement head. The backbone shares one scene space for understanding and future prediction, and the diffusion head maps its world tokens to future scene latents for dense decoding.

3 METHOD

3.1 WORLD-LANGUAGE-ACTION MODELS FOR DRIVING

Driving Formulation. Autonomous driving aims to plan a safe and feasible ego trajectory from past observations. We denote the visual history by $\mathcal{H}_t = \{I_{t-T+1}, \dots, I_t\}$, where $I_t = \{I_t^v\}_{v=1}^N$ contains synchronized images from N cameras. The planning task outputs a trajectory $\tau_{t+1} = (\xi_{t+1}, \dots, \xi_{t+K})$ starting at the next time step through

$$\hat{\tau}_{t+1} = f_{\text{drive}}(\mathcal{H}_t). \tag{1}$$

Ego states and past executed actions provide additional motion context when available. Scene understanding and future prediction support this planning objective by describing the current environment and the consequences of possible actions.

VLA and WAM Formulations. VLA models use visual history and language to predict a driving trajectory (Hwang et al., 2024; Jiang et al., 2024; Zhou et al., 2025b). WAMs connect future-scene prediction with trajectory planning from observed history and action context (Wang et al., 2024c; Zhang et al., 2025; Xia et al., 2026). We summarize their inputs and next-step outputs as

$$\hat{\tau}_{t+1} = f_{\text{VLA}}(\mathcal{H}_t, \ell_t), \tag{2}$$

$$(\hat{x}_{t+1}, \hat{\tau}_{t+1}) = f_{\text{WAM}}(\mathcal{H}_t, a_{\le t}), \tag{3}$$

where ℓ_t is language information, $a_{\le t}$ is the action history, and x_{t+1} is a future observation or scene state. They link planning to semantic understanding and future prediction, respectively.

Our WLA Framework. We propose a world-language-action framework to jointly model world states, language, and actions with a shared autoregressive backbone. We introduce an explicit world representation into this framework to describe the driving scene. This representation should jointly encode appearance, semantics, and geometry while integrating multiple camera views. It should provide a compact 3D scene state that preserves spatial relationships across views and supports temporal modeling. We treat this representation as a unified world modality and autoregressively

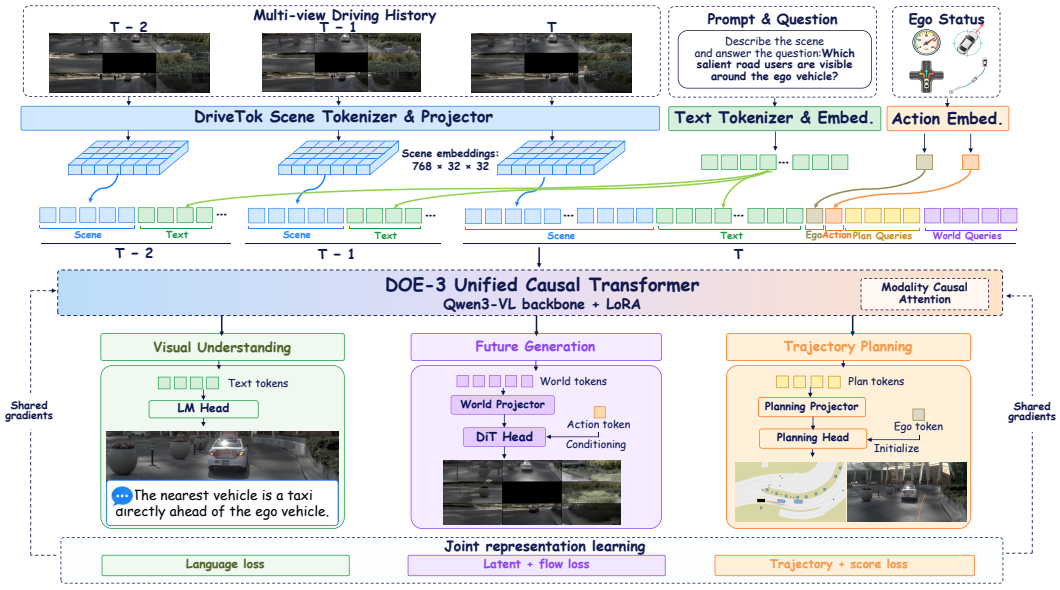


Figure 2: Overview of Doe-3. DriveTok encodes multi-view observations into scene latents, which interact with language, planning, action, and future-world tokens in a shared autoregressive backbone. Task-specific heads decode language, trajectories, and future scenes.

model its relationships with language and action. We encode multi-view observations as $z_t = \mathcal{E}(I_t)$ and express the framework as a mapping from multimodal history to next-step outputs

$$(\hat{z}_{t+1}, \hat{\ell}_{t+1}, \hat{\tau}_{t+1}) = f_{\text{WLA}}(z_{\leq t}, \ell_{\leq t}, a_{\leq t}). \tag{4}$$

The inputs contain world, language, and action history, while the outputs comprise the next world state, a language response, and a planned trajectory. Figure 2 presents the overall architecture. We use the same world representation for observed scenes and future prediction targets, allowing understanding and generation to supervise a common temporal model.

3.2 Unified Autoregressive Modeling

Driving World Representation. We instantiate the driving world representation with the pretrained DriveTok tokenizer (Zhuo et al., 2026b), which encodes synchronized multi-view images into a spatial grid $z_t \in \mathbb{R}^{64\times32\times32}$. DriveTok aggregates foundation-model features through 3D deformable attention and learns its representation with RGB, depth, semantic, and occupancy supervision. We project this grid into $HW = 1024$ scene tokens for the autoregressive backbone. We retain the pretrained representation and decoders as the basis for modeling world, language, and action.

Shared Autoregressive Backbone. We initialize the backbone from the text transformer of Qwen3-VL and adapt it with low-rank adaptation (Qwen Team, 2025). We construct the visual input directly from DriveTok scene tokens. For each frame, we arrange the modality tokens in the order

$$[s_t^{1:HW},\ d_t^{1:L_t},\ v_t^{1:Q_t},\ p_t^{1:R},\ a_t^{1:A},\ w_t^{1:HW}], \tag{5}$$

where s_t represents the observed scene, d_t and v_t are description and question–answer text tokens, p_t supports planning, a_t encodes action u_t and ego state, and w_t supports future-world prediction. We use $R = 32$ planning tokens, and text lengths depend on the frame and task. We concatenate these sequences over time and apply causal attention within and across frames. Multi-axis rotary positions preserve temporal and spatial coordinates for scene and world tokens.

Causal Dependencies. Text tokens interact directly with the observed scene tokens in the shared backbone. Planning tokens read the observed history and available description tokens without accessing the current supervised action. We place world tokens after the action token to condition future prediction on that action. During training, we provide the executed action from t to $t+1$ at this position, while the next ground-truth scene remains in the causal future. The current VQA answer block is masked from planning and world tokens. This ordering separates the information available for planning from the additional action condition used for world prediction.

Visual Decoding. We introduce a lightweight diffusion head to map autoregressive world tokens to a future latent in the DriveTok representation space. We project the world tokens into dense conditions for a compact diffusion transformer (Peebles & Xie, 2023). A residual action projection supplies action information, and a local spatial mixer connects neighboring positions in the scene grid. The diffusion head predicts $\hat{z}_{t+1}$ from noise under these conditions. We then use the pretrained DriveTok decoders to recover dense scene outputs

$$\hat{o}^k_{t+1} = \mathcal{D}^k_{\mathrm{DT}}(\hat{z}_{t+1}), \qquad k \in \{\mathrm{RGB}, \mathrm{depth}, \mathrm{semantic}, \mathrm{occupancy}\}. \tag{6}$$

This head keeps generation in the scene space the backbone reads for understanding and planning.

Language Decoding. We apply the pretrained language head to the text hidden states from the shared backbone. Conditioned on the scene tokens and a language prompt, the model autoregressively decodes a scene description or a VQA answer. During training, we supervise description tokens with next-token prediction and VQA responses with answer-token prediction in the shared sequence. Language supervision updates the backbone through scene-text interactions.

Action Decoding. We project the planning tokens to 32 planning registers and decode trajectories with a proposal-and-score planning head. The head predicts 64 trajectory proposals with eight future poses each, together with proposal-conditioned dynamics and six NAVSIM sub-scores. We use the predicted scores to select the output trajectory. The planning head reads the causal history before the current action and future-world tokens, so it produces a trajectory from observed context.

3.3 Training

Training Objectives. We jointly optimize planning, future-scene prediction, and grounded language generation through

$$\mathcal{L} = \mathcal{L}_{\mathrm{plan}} + \mathcal{L}_{\mathrm{gen}} + \mathcal{L}_{\mathrm{lang}}, \tag{7}$$

where each term includes its corresponding loss weights. The planning objective supervises trajectories, proposal-conditioned dynamics, and proposal scores. The generation objective combines latent matching and flow matching to supervise future DriveTok latents and the autoregressive world tokens through the diffusion head. The language objective supervises scene descriptions and VQA answers through the language head. All three objectives update the shared backbone through their respective modality tokens. We provide the loss decomposition and weights in Section A.

Training Stages. We first align the interface between DriveTok and the backbone while freezing the backbone, task heads, and visual stack. We then train only the LoRA parameters of the backbone while keeping the interface and task heads frozen. We retain gradients through the frozen heads so that planning, generation, and language supervision can adapt the backbone. Finally, we jointly tune LoRA, the scene interface, the planning head, and the conditioning modules of the diffusion head with decreasing learning-rate multipliers. This schedule progresses from aligning the scene input to adapting the shared temporal model and jointly optimizing its task outputs.

4 Experiments

4.1 Experimental Setup

Benchmarks and Metrics. We evaluate trajectory planning on NAVSIM v1 (Dauner et al., 2024) and NAVSIM v2 (Cao et al., 2025) and future scene generation on nuScenes (Caesar et al., 2020). Planning tables report PDMS for v1 and EPDMS for v2 together with their component metrics. Future image generation is evaluated with FID. Depth generation is evaluated with absolute relative error and threshold accuracies at 1.25 and 1.25^2. Occupancy generation is evaluated with semantic mIoU and geometric IoU at the forecast horizons listed in Table 5. These measurements assess driving behavior, visual appearance, and future scene geometry.

Implementation. We initialize the backbone from Qwen3-VL (Qwen Team, 2025) with LoRA and feed it all 1024 DriveTok scene tokens of each observed frame. We train on OpenScene (OpenScene Contributors, 2023) and nuScenes (Caesar et al., 2020) with the three-stage schedule, using AdamW (Loshchilov & Hutter, 2017) and a cosine learning-rate schedule on 64 NVIDIA A800 GPUs. Per-stage epochs, learning rates, and optimizer settings are listed in Section A.

Table 1: Planning results on NAVSIM v1 `navtest`. C and L denote camera and LiDAR inputs. The shaded final column reports PDMS, and all metrics are higher-is-better.

Method	Input	NC↑	DAC↑	TTC↑	Comf.↑	EP↑	PDMS↑
Perception-based Methods							
VADv2 (Chen et al., 2024)	C	97.2	89.1	91.6	100	76.0	80.9
UniAD (Hu et al., 2023b)	C	97.8	91.9	92.9	100	78.8	83.4
Transfuser (Chitta et al., 2022)	C & L	97.7	92.8	92.8	100	79.2	84.0
DiffusionDrive (Liao et al., 2025a)	C & L	98.2	96.2	94.7	100	82.2	88.1
WoTE (Li et al., 2025g)	C & L	98.5	96.8	94.9	99.9	81.9	88.3
DriveSuprim (Yao et al., 2025)	C & L	97.8	97.3	93.6	100	86.7	89.9
VLA-based Methods							
AdaThinkDrive (Luo et al., 2025)	C	98.5	94.4	94.9	100	79.0	86.2
AutoVLA (Zhou et al., 2025c)	C	98.4	95.6	98.0	99.9	81.9	89.1
DriveDreamer-Policy (Zhou et al., 2026b)	C	98.4	97.1	95.1	100	83.5	89.2
ReCogDrive (Li et al., 2025h)	C	98.2	97.8	95.2	99.8	83.5	89.6
DriveVLA-W0 (Li et al., 2025f)	C	98.7	99.1	95.3	99.3	83.3	90.2
ExploreVLA (Sheng et al., 2026)	C	98.8	98.4	96.5	99.9	83.5	90.4
Qwen-Drive-1.0-RL (Zhou et al., 2026a)	C	98.6	98.2	95.9	100	84.8	90.7
WAM-based Methods							
LAW (Li et al., 2024)	C	96.4	95.4	88.7	99.9	81.7	84.6
Epona (Zhang et al., 2025)	C	97.9	95.1	93.8	99.9	80.4	86.2
PWM (Zhao et al., 2025)	C	98.6	95.9	95.4	100	81.8	88.1
DriveLaW (Xia et al., 2026)	C	99.0	97.1	96.7	100	81.3	89.1
DriveWAM (Shi et al., 2026)	C	98.3	98.1	95.2	100	84.3	90.1
EponaV2 (Xu et al., 2026)	C	98.6	97.9	95.7	100	84.8	90.4
DriveVA (Liu et al., 2026)	C	99.2	97.5	98.7	100	83.5	90.9
World-Language-Action Model							
Doe-3	C	98.8	97.5	96.3	100	93.2	**94.0**

Table 2: Planning results on NAVSIM v2 `navtest`. All scores follow the v2 evaluation protocol.

Method	NC↑	DAC↑	DDC↑	TL↑	EP↑	TTC↑	LK↑	HC↑	EC↑	EPDMS↑
Ego Status MLP	93.1	77.9	92.7	99.6	86.0	91.5	89.4	98.3	85.4	64.0
Transfuser (Chitta et al., 2022)	96.9	89.9	97.8	99.7	87.1	95.4	92.7	98.3	87.2	76.7
Hydra-MDP++ (Li et al., 2025d)	97.2	97.5	99.4	99.6	83.1	96.5	94.4	98.2	70.9	81.4
DriveSuprim (Yao et al., 2025)	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
DiffusionDrive (Liao et al., 2025a)	98.2	95.9	99.4	99.8	87.5	97.3	96.8	98.3	87.7	84.5
DriveVLA-W0 (Li et al., 2025f)	98.5	99.1	98.0	99.7	86.4	98.1	93.2	97.9	58.9	86.1
DriveLaW (Xia et al., 2026)	98.7	96.9	99.6	99.8	87.5	98.3	97.6	98.4	77.4	88.6
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
EponaV2 (Xu et al., 2026)	98.5	97.4	99.5	99.9	87.9	98.1	97.7	98.2	77.4	88.9
Doe-3	98.8	98.8	98.9	99.4	91.0	98.3	95.7	98.2	85.2	**90.8**

4.2 Trajectory Planning on NAVSIM

Tables 1 and 2 compare Doe-3 with existing driving models. The v1 table groups methods by their modeling approach and reports the input modality. The component scores characterize collision avoidance, drivable-area compliance, time-to-collision, comfort, and progress, alongside the aggregate PDMS. The v2 table adds the extra v2 driving criteria and aggregate EPDMS. Doe-3 achieves a PDMS of 94.0 on NAVSIM v1 and an EPDMS of 90.8 on NAVSIM v2. On v1, Doe-3 obtains an EP score of 93.2 and NC, DAC, and TTC scores of 98.8, 97.5, and 95.9.

4.3 Multi-Modal Future Generation on nuScenes

Future Images. Table 3 evaluates future RGB generation. Doe-3 generates multi-view images by decoding predicted scene latents, using the observed history and ego-action sequence as context. Its FID is 9.91. Figure 4 visualizes rollouts under different action conditions from the same observed scene, relating action control to changes in the road layout and surrounding views.

Table 3: Future image generation on nuScenes. We report FID and mark pretrained models, volumetric conditions, and multi-view output. Bold and underline mark the top two per group.

Method	Pre-trained Model-Free	Volumetric Condition-Free	Multi-view	FID ↓
WoVoGen (Lu et al., 2023)	✗	✗	✓	27.60
X-Scene (Yang et al., 2025)	✗	✗	✓	<u>11.29</u>
OccScene (Li et al., 2025b)	✗	✗	✓	11.87
UniScene (Li et al., 2025a)	✗	✗	✓	**6.45**
DriveDreamer (Wang et al., 2024b)	✗	✓	✗	14.90
DrivingGPT (Chen et al., 2025)	✗	✓	✗	<u>12.78</u>
Vista (Gao et al., 2024b)	✗	✓	✗	**6.90**
Doe-1 (Zheng et al., 2024b)	✗	✓	✗	15.90
MagicDrive (Gao et al., 2023)	✗	✓	✓	16.20
Panacea (Wen et al., 2024)	✗	✓	✓	16.96
Drive-WM (Wang et al., 2024c)	✗	✓	✓	15.80
DriveDreamer-2 (Zhao et al., 2024)	✗	✓	✓	<u>11.20</u>
DiST-4D (Guo et al., 2025)	✓	✗	✓	**7.40**
DriveGAN (Kim et al., 2021)	✓	✓	✗	73.40
DrivingWorld (Hu et al., 2024)	✓	✓	✗	**7.40**
Epona (Zhang et al., 2025)	✓	✓	✗	<u>7.50</u>
BEVGen (Swerdlow et al., 2024)	✓	✓	✓	25.54
GenAD (Zheng et al., 2024a)	✓	✓	✓	**15.40**
MagicDrive-V2 (Gao et al., 2024a)	✓	✓	✓	<u>20.91</u>
Doe-2 (Zhuo et al., 2026a)	✓	✓	✓	10.48
Doe-3	✗	✓	✓	**9.91**

Table 4: Future depth generation on nuScenes. We report AbsRel and threshold accuracies.

Method	Abs. Rel. ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$
SurroundDepth (Wei et al., 2023)	0.28	0.66	0.84
M^2Depth (Zou et al., 2024)	0.26	0.73	0.87
Dist-4D (Guo et al., 2025)	0.39	0.58	0.81
OmniNWM (Li et al., 2025c)	<u>0.23</u>	<u>0.81</u>	<u>0.93</u>
Doe-3	**0.13**	**0.89**	**0.94**

Future Depth. Table 4 evaluates the metric geometry of the generated future scene. Doe-3 obtains an absolute relative error of 0.13, with threshold accuracies of 0.89 and 0.94. These results indicate accurate metric geometry in the predicted scene latents.

Future Occupancy. Table 5 reports semantic mIoU and geometric IoU across future horizons. Doe-3 attains average scores of 27.03 and 61.49, respectively. The horizon-wise results characterize the evolution of semantic and geometric accuracy during rollout. Occupancy and depth complement image generation by evaluating spatial structure in the predicted world state.

4.4 Experimental Analysis

Planning Ablation. We compare three variants: a perception-based model (P), a generation-based model (P+G), and the generation-language model (P+G+L). The perception-based variant predicts trajectories from observed scene features. The generation-based variant adds future-scene modeling. The generation-language variant further incorporates language supervision and scene–text interaction. Table 6 reports their planning results, and Figure 3 compares their trajectories.

The P-to-P+G comparison measures the contribution of future generation, while the P+G-to-P+G+L comparison measures the additional contribution of language. The corresponding PDMS differences are 2.5 and 4.0, and the EPDMS differences are 2.6 and 2.0.

Where Language Changes Planning. We align the three variants by scenario token over the same 12,146 valid NAVSIM-v1 scenes. Doe-3 (P+G+L) improves PDMS by 3.654 percentage points over P+G and by 6.459 points over P. The largest component gain is ego progress: +6.374 and +11.669 points, respectively. TTC improves by 1.877/1.861 points and no-at-fault collision scores

Table 5: Quantitative results for 4D occupancy forecasting on Occ3D-nuScenes. Avg. averages 1s, 2s, and 3s. OccWorld-D uses predictions from TPVFormer (Huang et al., 2023). "-STC" and "-F" use occupancy from STCOcc (Liao et al., 2025b) and FB-Occ (Li et al., 2023).

Method	Input	Semantic mIoU (%) ↑				Geometric IoU (%) ↑			
		1s	2s	3s	Avg.	1s	2s	3s	Avg.
RenderWorld (Yan et al., 2025)	Occ GT	28.69	18.89	14.83	20.89	37.74	28.41	24.08	30.08
Occ-LLM (Xu et al., 2025b)	Occ GT	24.02	21.65	17.29	20.99	36.65	32.14	28.77	35.52
DFIT-OccWorld (Zhang et al., 2024)	Occ GT	31.68	21.29	15.19	22.71	40.28	31.24	25.29	32.27
DOME (Gu et al., 2024)	Occ GT	35.11	25.89	20.29	27.10	43.99	35.36	29.74	36.36
UniScene (Li et al., 2025a)	Occ GT	35.37	29.59	25.08	31.76	38.34	32.70	29.09	34.84
I^2-World (Liao et al., 2025c)	Occ GT	**47.62**	**38.58**	**32.98**	**39.73**	**54.29**	**49.43**	**45.69**	**49.80**
PreWorld (Li et al., 2025e)	Camera	12.27	9.24	7.15	9.55	23.62	21.62	19.63	21.62
Occ-LLM (Xu et al., 2025b)	Camera	11.28	10.21	9.13	10.21	27.11	24.07	20.19	23.79
DFIT-OccWorld (Zhang et al., 2024)	Camera	13.38	10.16	7.96	10.50	19.18	16.85	25.02	17.02
DOME-STC (Gu et al., 2024)	Camera	17.79	14.23	11.58	14.53	26.39	23.20	20.42	23.33
I^2-World-STC (Liao et al., 2025c)	Camera	21.67	18.78	16.47	18.97	30.55	28.76	26.99	28.77
DOME-F (Gu et al., 2024)	Camera	24.12	17.41	13.24	18.25	35.18	27.90	23.44	28.84
DTT (Xu et al., 2025a)	Camera	24.87	18.30	15.63	19.60	38.98	37.45	31.89	36.11
COME-F (Shi et al., 2025)	Camera	26.56	21.73	18.49	22.26	48.08	43.84	40.28	44.07
Doe-3	Camera	**31.58**	**26.44**	**23.06**	**27.03**	**65.11**	**61.27**	**58.10**	**61.49**

Table 6: Contributions of generation and language to planning. G and L indicate future-generation and language objectives. All three variants perform trajectory planning.

Model	G	L	NC↑	DAC↑	TTC↑	Comf.↑	EP↑	v1 PDMS↑	v2 EPDMS↑
Perception-based (P)	×	×	97.8	96.4	94.1	100.0	81.6	87.5	86.2
Generation-based (P+G)	✓	×	98.0	97.9	93.3	100.0	86.5	90.0	88.8
Generation-language (P+G+L)	✓	✓	98.8	97.5	95.9	100.0	93.2	94.0	90.8

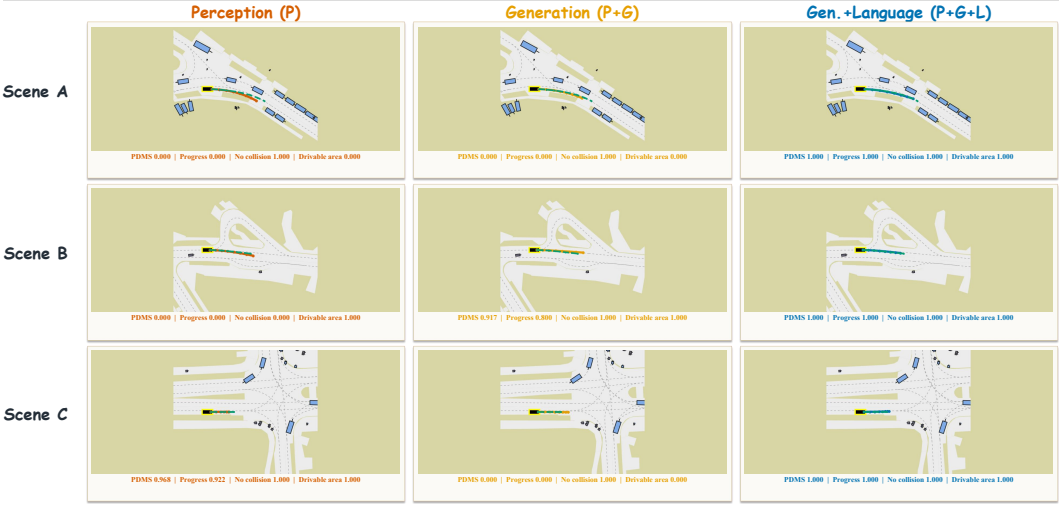


Figure 3: Planning comparison. Planning trajectories of perception (P), generation-augmented (P+G), and world-language-action-augmented (P+G+L) models in the same scenes. The green dashed line is the ground-truth trajectory.

by 0.659/1.009 points, while comfort remains nearly unchanged. These results associate the joint world–language–action representation with more effective route progress alongside improved interaction timing and collision avoidance. The gains are selective: relative to P+G, drivable-area compliance increases by only 0.107 points, while driving-direction compliance decreases by 0.762 points. Figure 3 compares the three variants in a common scene coordinate frame.

Generation Ablation. Table 7 compares P+G and P+G+L on the same image, depth, and occupancy metrics. This comparison connects language supervision to both the appearance and geometry of the predicted scene. The P+G+L model reduces FID from 10.48 to 9.91 and absolute relative depth error from 0.15 to 0.13. Depth accuracy at $\delta < 1.25$ increases from 0.87 to 0.89, while accuracy at $\delta < 1.25^2$ remains 0.94. Average occupancy mIoU and IoU increase from 26.35 to 27.03 and from 60.79 to 61.49, respectively. These modality-specific changes identify which aspects of future prediction benefit most from the shared scene–text representation.

Table 7: Contribution of language to future generation on nuScenes. The generation and generation-language variants are compared across RGB, depth, and occupancy. δ_1 and δ_2 use thresholds of 1.25 and 1.25^2, and occupancy metrics are arithmetic means over 1, 2, and 3 seconds.

Model	RGB	Depth			Occupancy	
	FID↓	Abs. Rel.↓	δ_1↑	δ_2↑	mIoU↑	IoU↑
Generation-based (P+G)	10.48	0.15	0.87	**0.94**	26.35	60.79
Generation-language (P+G+L)	**9.91**	**0.13**	**0.89**	**0.94**	**27.03**	**61.46**

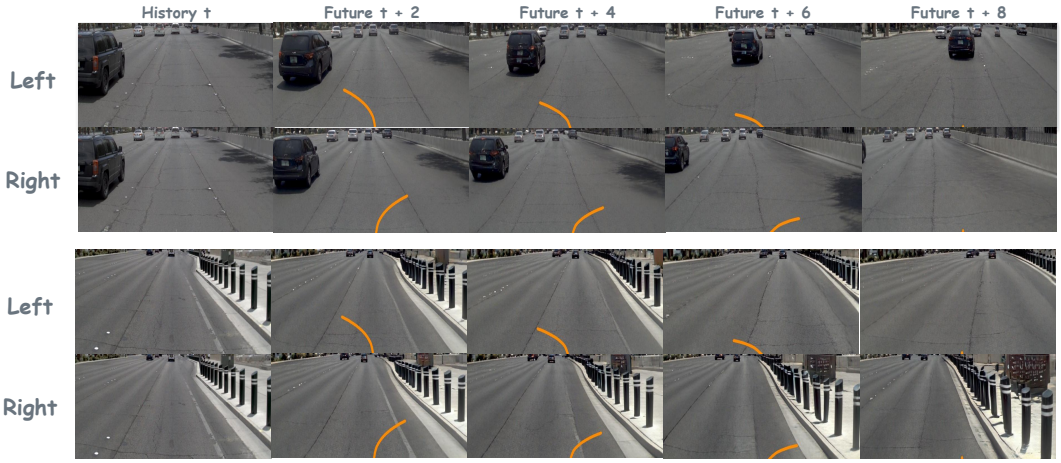


Figure 4: Controllable future generation. Doe-3 generates future images under different steering conditions, with the corresponding ego trajectories drawn in orange.

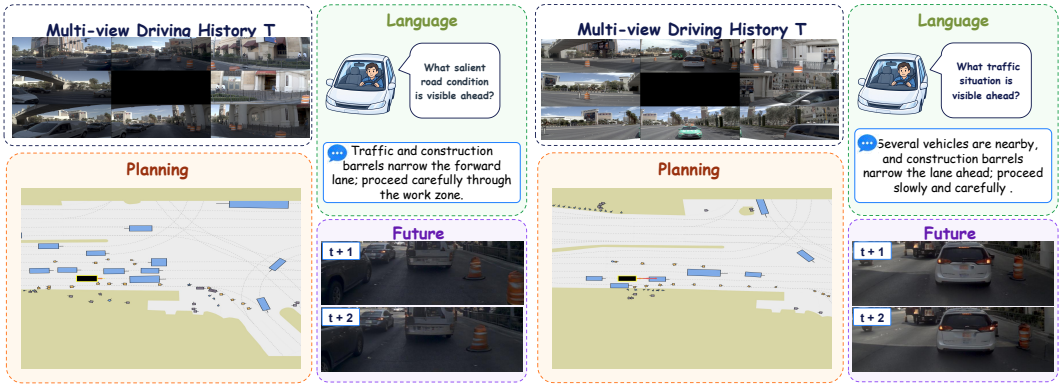


Figure 5: Joint language, planning, and generation. Outputs use one multi-view history.

Action-Controlled Futures. Figure 4 holds the observation history fixed and varies the future ego-action sequence. Each row follows one action condition through time. The layout exposes the relation between the commanded motion and the resulting view evolution, including changes in road position and the relative placement of nearby objects.

Language, Planning, and Generation in One Scene. Figure 5 presents a scene description or answer, a planned trajectory, and future scene predictions for the same input history. The shared presentation relates the objects and spatial relationships expressed in language to the driving context and predicted future. This makes the three outputs interpretable within a common scene.

5 Conclusion

We presented Doe-3, a world-language-action model that connects understanding, prediction, and action through a shared spatial scene representation. Text tokens directly attend to the full surround-view scene history, while action-conditioned future prediction models scene evolution in the same latent space. Language, generation, and planning objectives jointly shape the backbone, connecting semantic context and predictive dynamics to driving decisions. A lightweight diffusion head supports future RGB, depth, and occupancy decoding alongside trajectory planning. Evaluation on NAVSIM v1/v2 and nuScenes, together with perception, generation, and generation-language comparisons, examines these capabilities and the contributions of the training objectives.

AI USE STATEMENT

In this work, we have not used generative AI tools for any tasks with required disclosure, and all such tasks are not applicable to this work. We used generative AI tools only to polish the writing of parts of the manuscript. The authors reviewed all AI-assisted edits to ensure that they preserve the intended technical meaning and contain no unsupported claims. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, the main paper and Section A provide detailed descriptions of our methodology and implementation. We will release the complete source code and model weights.

REFERENCES

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

Wei Cao, Marcel Hallgarten, Tianyu Li, Daniel Dauner, Xunjiang Gu, Caojun Wang, Yakov Miron, Marco Aiello, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. Pseudo-simulation for autonomous driving. In *Conference on Robot Learning*, 2025.

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. *arXiv preprint arXiv:2402.13243*, 2024. URL https://arxiv.org/abs/2402.13243.

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.

Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. TransFuser: Imitation with Transformer-Based Sensor Fusion for Autonomous Driving. *arXiv preprint arXiv:2205.15997*, 2022. URL https://arxiv.org/abs/2205.15997.

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

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. *arXiv preprint arXiv:2504.19580*, 2025. URL https://arxiv.org/abs/2504.19580.

Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. *arXiv preprint arXiv:2310.02601*, 2023.

Ruiyuan Gao, Kai Chen, Bo Xiao, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive-v2: High-resolution long video generation for autonomous driving with adaptive control. *arXiv preprint arXiv:2411.13807*, 2024a.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems*, 37:91560–91596, 2024b.

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.

Jiazhe Guo, Yikang Ding, Xiwu Chen, Shuo Chen, Bohan Li, Yingshuang Zou, Xiaoyang Lyu, Feiyang Tan, Xiaojuan Qi, Zhiheng Li, et al. Dist-4d: Disentangled spatiotemporal diffusion with metric depth for 4d driving scene generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27231–27241, 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.

Xiaotao Hu, Wei Yin, Mingkai Jia, Junyuan Deng, Xiaoyang Guo, Qian Zhang, Xiaoxiao Long, and Ping Tan. Drivingworld: Constructing world model for autonomous driving via video gpt. *arXiv preprint arXiv:2412.19505*, 2024.

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. *arXiv preprint arXiv:2212.10156*, 2023b.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9223–9232, 2023.

Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, Yin Zhou, James Guo, Dragomir Anguelov, and Mingxing Tan. 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. *arXiv preprint arXiv:2303.12077*, 2023.

Bo Jiang, Shaoyu Chen, Bencheng Liao, Xingyu Zhang, Wei Yin, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Senna: Bridging large vision-language models and end-to-end autonomous driving. *arXiv preprint arXiv:2410.22313*, 2024.

Seung Wook Kim, Jonah Philion, Antonio Torralba, and Sanja Fidler. Drivegan: Towards a controllable high-quality neural simulation. In *CVPR*, 2021.

Bohan Li, Jiazhe Guo, Hongsi Liu, Yingshuang Zou, Yikang Ding, Xiwu Chen, Hu Zhu, Feiyang Tan, Chi Zhang, Tiancai Wang, et al. Uniscene: Unified occupancy-centric driving scene generation. In *Proceedings of the computer vision and pattern recognition conference*, pp. 11971–11981, 2025a.

Bohan Li, Xin Jin, Jianan Wang, Yukai Shi, Yasheng Sun, Xiaofeng Wang, Zhuang Ma, Baao Xie, Chao Ma, Xiaokang Yang, et al. Occscene: Semantic occupancy-based cross-task mutual learning for 3d scene generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b.

Bohan Li, Zhuang Ma, Dalong Du, Baorui Peng, Zhujin Liang, Zhenqiang Liu, Chao Ma, Yueming Jin, Hao Zhao, Wenjun Zeng, et al. Omninwm: Omniscient driving navigation world models. *arXiv preprint arXiv:2510.18313*, 2025c.

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*, 2025d. URL https://arxiv.org/abs/2503.12820.

Xiang Li, Pengfei Li, Yupeng Zheng, Wei Sun, Yan Wang, and Yilun Chen. Semi-supervised vision-centric 3d occupancy world model for autonomous driving. *arXiv preprint arXiv:2502.07309*, 2025e.

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. *arXiv preprint arXiv:2406.08481*, 2024. URL https://arxiv.org/abs/2406.08481.

Yingyan Li, Shuyao Shang, Weisong Liu, Bing Zhan, Haochen Wang, Yuqi Wang, Yuntao Chen, Xiaoman Wang, Yasong An, Chufeng Tang, et al. Drivevla-w0: World models amplify data scaling law in autonomous driving. *arXiv preprint arXiv:2510.12796*, 2025f.

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. *arXiv preprint arXiv:2504.01941*, 2025g. URL https://arxiv.org/abs/2504.01941.

Yongkang Li, Kaixin Xiong, Xiangyu Guo, Fang Li, Sixu Yan, Gangwei Xu, Lijun Zhou, Long Chen, Haiyang Sun, Bing Wang, et al. ReCogDrive: A Reinforced Cognitive Framework for End-to-End Autonomous Driving. *arXiv preprint arXiv:2506.08052*, 2025h. URL https://arxiv.org/abs/2506.08052.

Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. Fb-occ: 3d occupancy prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492*, 2023.

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, 2025a.

Zhimin Liao, Ping Wei, Shuaijia Chen, Haoxuan Wang, and Ziyang Ren. Stcocc: Sparse spatial-temporal cascade renovation for 3d occupancy and scene flow prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1516–1526, 2025b.

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*, pp. 25810–25819, 2025c.

Mengmeng Liu, Diankun Zhang, Jiuming Liu, Jianfeng Cui, Hongwei Xie, Guang Chen, Hangjun Ye, Michael Ying Yang, Francesco Nex, and Hao Cheng. DriveVA: Video Action Models are Zero-Shot Drivers. *arXiv preprint arXiv:2604.04198*, 2026. URL https://arxiv.org/abs/2604.04198.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jiachen Lu, Ze Huang, Jiahui Zhang, Zeyu Yang, and Li Zhang. Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. *arXiv preprint arXiv:2312.02934*, 2023.

Yuechen Luo, Fang Li, Shaoqing Xu, Zhiyi Lai, Lei Yang, Qimao Chen, Ziang Luo, Zixun Xie, Shengyin Jiang, Jiaxin Liu, Long Chen, Bing Wang, and Zhi xin Yang. AdaThinkDrive: Adaptive Thinking via Reinforcement Learning for Autonomous Driving. *arXiv preprint arXiv:2509.13769*, 2025. URL https://arxiv.org/abs/2509.13769.

Alexander Naumann, Xunjiang Gu, Tolga Dimlioglu, Mariusz Bojarski, Alperen Degirmenci, Alexander Popov, Devansh Bisla, Marco Pavone, Urs Müller, and Boris Ivanovic. Data scaling laws for end-to-end autonomous driving. *arXiv preprint arXiv:2504.04338*, 2025.

NVIDIA. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

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

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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. URL https://arxiv.org/abs/2604.02714.

Chen Shi, Jinrui Xu, Shaoshuai Shi, Kehua Sheng, Bo Zhang, and Li Jiang. DriveWAM: Video Generative Priors Enable Scalable World-Action Modeling for Autonomous Driving. *arXiv preprint arXiv:2605.28544*, 2026. URL https://arxiv.org/abs/2605.28544.

Yining Shi, Kun Jiang, Qiang Meng, Ke Wang, Jiabao Wang, Wenchao Sun, Tuopu Wen, Mengmeng Yang, and Diange Yang. Come: Adding scene-centric forecasting control to occupancy world model. *arXiv preprint arXiv:2506.13260*, 2025.

Alexander Swerdlow, Runsheng Xu, and Bolei Zhou. Street-view image generation from a bird's-eye view layout. *IEEE Robotics and Automation Letters*, 2024.

Shuhan Tan, Kashyap Chitta, Yuxiao Chen, Ran Tian, Yurong You, Yan Wang, Wenjie Luo, Yulong Cao, Philipp Krähenbühl, Marco Pavone, and Boris Ivanovic. Latent chain-of-thought world modeling for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39724–39733, 2026.

Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M. Alvarez. OmniDrive: A holistic LLM-agent framework for autonomous driving with 3d perception, reasoning and planning. *arXiv preprint arXiv:2405.01533*, 2024a.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drive-Dreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 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, 2024b.

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*, pp. 14749–14759, 2024c.

Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Yongming Rao, Guan Huang, Jiwen Lu, and Jie Zhou. Surrounddepth: Entangling surrounding views for self-supervised multi-camera depth estimation. In *Conference on robot learning*, pp. 539–549. PMLR, 2023.

Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea: Panoramic and controllable video generation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6902–6912, 2024.

Tianze Xia, Yongkang Li, Lijun Zhou, Jingfeng Yao, Kaixin Xiong, Haiyang Sun, Bing Wang, Kun Ma, Hangjun Ye, Wenyu Liu, 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*, 2026.

Zhexiao Xiong, Xin Ye, Burhan Yaman, Sheng Cheng, Yiren Lu, Jingru Luo, Nathan Jacobs, and Liu Ren. UniDrive-WM: Unified understanding, planning and generation world model for autonomous driving. *arXiv preprint arXiv:2601.04453*, 2026.

Haoran Xu, Peixi Peng, Guang Tan, Yiqian Chang, Yisen Zhao, and Yonghong Tian. Delta-triplane transformers as occupancy world models. *arXiv preprint arXiv:2503.07338*, 2025a.

Jiawei Xu, Zhizhou Zhong, Zhijian Shu, Mingkai Jia, Mingxiao Li, Jia-Wang Bian, Qian Zhang, Kaicheng Zhang, Jin Xie, Jian Yang, and Wei Yin. EponaV2: Driving World Model with Comprehensive Future Reasoning. *arXiv preprint arXiv:2605.14696*, 2026. URL https://arxiv.org/abs/2605.14696.

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)*, pp. 8434–8441. IEEE, 2025b.

Ziyang Yan, Wenzhen Dong, Yihua Shao, Yuhang Lu, Haiyang Liu, Jingwen Liu, Haozhe Wang, Zhe Wang, Yan Wang, Fabio Remondino, et al. Renderworld: World model with self-supervised 3d label. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6063–6070. IEEE, 2025.

Yi Yang, Zhihong Liu, Siqi Kou, Yiyang Chen, Yanzhe Hu, Jianbo Zhou, Boyuan Zhao, Zhijie Wei, Xiao Xia, Xueqi Li, Pengfei Liu, and Zhijie Deng. World-language-action model for unified world modeling, language reasoning, and action synthesis. *arXiv preprint arXiv:2606.05979*, 2026.

Yu Yang, Alan Liang, Jianbiao Mei, Yukai Ma, Yong Liu, and Gim Hee Lee. X-scene: Large-scale driving scene generation with high fidelity and flexible controllability. *arXiv preprint arXiv:2506.13558*, 2025.

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. *arXiv preprint arXiv:2506.06659*, 2025. URL https://arxiv.org/abs/2506.06659.

Haiming Zhang, Ying Xue, Xu Yan, Jiacheng Zhang, Weichao Qiu, Dongfeng Bai, Bingbing Liu, Shuguang Cui, and Zhen Li. An efficient occupancy world model via decoupled dynamic flow and image-assisted training. *arXiv preprint arXiv:2412.13772*, 2024.

Kaiwen Zhang, Zhenyu Tang, Xiaotao Hu, Xingang Pan, Xiaoyang Guo, Yuan Liu, Jingwei Huang, Li Yuan, Qian Zhang, Xiao-Xiao Long, Xun Cao, and Wei Yin. Epona: Autoregressive diffusion world model for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27220–27230, 2025.

Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. *arXiv preprint arXiv:2403.06845*, 2024.

Zhida Zhao, Talas Fu, Yifan Wang, Lijun Wang, and Huchuan Lu. From Forecasting to Planning: Policy World Model for Collaborative State-Action Prediction. *arXiv preprint arXiv:2510.19654*, 2025. URL https://arxiv.org/abs/2510.19654.

Wenzhao Zheng, Ruiqi Song, Xianda Guo, Chenming Zhang, and Long Chen. Genad: Generative end-to-end autonomous driving. In *European Conference on Computer Vision*, pp. 87–104. Springer, 2024a.

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*, 2024b.

Xin Zhou, Dingkang Liang, Sifan Tu, Xiwu Chen, Yikang Ding, Dingyuan Zhang, Feiyang Tan, Hengshuang Zhao, and Xiang Bai. HERMES: A unified self-driving world model for simultaneous 3d scene understanding and generation. *arXiv preprint arXiv:2501.14729*, 2025a.

Xin Zhou, Zongchuang Zhao, Zhibo Yang, Mingsheng Li, Humen Zhong, Shuai Bai, Du Chu, Ruizhe Chen, Zhaohai Li, Jun Tang, Qiuyue Wang, Mingkun Yang, Jiazhao Zhang, Dayiheng Liu, Dingkang Liang, and Xiang Bai. Qwen-Drive-1.0: An Initial Step towards a Vision-Language Foundation Model for Autonomous Driving. *arXiv preprint arXiv:2609.00111*, 2026a. URL https://arxiv.org/abs/2609.00111.

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

Yang Zhou, Xiaofeng Wang, Hao Shao, Letian Wang, Guosheng Zhao, Jiangnan Shao, Jiagang Zhu, Tingdong Yu, Zheng Zhu, Guan Huang, and Steven L. Waslander. DriveDreamer-Policy: A Geometry-Grounded World-Action Model for Unified Generation and Planning. *arXiv preprint arXiv:2604.01765*, 2026b. URL https://arxiv.org/abs/2604.01765.

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*, 2025c. URL https://arxiv.org/abs/2506.13757.

Dong Zhuo, Wenzhao Zheng, Sicheng Zuo, Yuanhui Huang, Siming Yan, Lu Hou, Jie Zhou, and Jiwen Lu. Doe-2: 3d representation world model for unified driving scene forecasting. In *European Conference on Computer Vision*, pp. 499–520. Springer, 2026a.

Dong Zhuo, Wenzhao Zheng, Sicheng Zuo, Siming Yan, Lu Hou, Jie Zhou, and Jiwen Lu. DriveTok: 3d driving scene tokenization for unified multi-view reconstruction and understanding. *arXiv preprint arXiv:2603.19219*, 2026b.

Yingshuang Zou, Yikang Ding, Xi Qiu, Haoqian Wang, and Haotian Zhang. M 2 depth: Self-supervised two-frame m ulti-camera m etric depth estimation. In *European Conference on Computer Vision*, pp. 269–285. Springer, 2024.

Sicheng Zuo, Yuxuan Li, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Vega: Learning to drive with natural language instructions. *arXiv preprint arXiv:2603.25741*, 2026.

Appendix

A Training and Evaluation Details

Scene Sequence. Each observed frame contributes 1024 spatial scene tokens. Text blocks have variable length, and planning and world queries provide task readouts. The sequence length is therefore determined by the observation window, text lengths, and query configuration. All observed spatial locations remain available to text attention.

Training Stages. The first stage aligns the scene interface to the pretrained backbone and task heads. The second stage adapts the shared model with language, planning, and future-generation supervision. The final planning-oriented stage uses planning and future-generation objectives. Training uses OpenScene and nuScenes on 64 A800 GPUs. Table 8 summarizes the schedules.

Table 8: Three-stage training configuration. All stages use AdamW on 64 NVIDIA A800 GPUs.

Configuration	Stage 1	Stage 2	Stage 3
OpenScene epochs	1	2	5
nuScenes epochs	40	40	80
Base learning rate	10^{-4}	5×10^{-5}	5×10^{-5}

Optimization. AdamW uses $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and weight decay 0.01. The learning rate follows a cosine schedule with a minimum of 10^{-6}. The base learning rates are 10^{-4} for stage 1 and 5×10^{-5} for stages 2 and 3, as listed in Table 8.

Planning is evaluated on the NAVSIM-v1 and NAVSIM-v2 `navtest` splits using PDMS and EPDMS, respectively, together with their component metrics. The NAVSIM-v1 comparison uses the same 12,146 valid scenarios for all three model variants. On nuScenes, future images are generated at a resolution of 256×704 and evaluated using FID. Depth generation is evaluated using absolute relative error (Abs Rel) and threshold accuracies at $\delta < 1.25$ and $\delta < 1.25^2$. Occupancy forecasting is evaluated using semantic mIoU and geometric IoU at prediction horizons of 1, 2, and 3 seconds, with average scores computed across these horizons.

B Geometric Detail.

Figure 6 shows predicted depth at $t + 1$ and $t + 2$ alongside the corresponding ground-truth multi-view images, visualizing road geometry and surrounding objects.

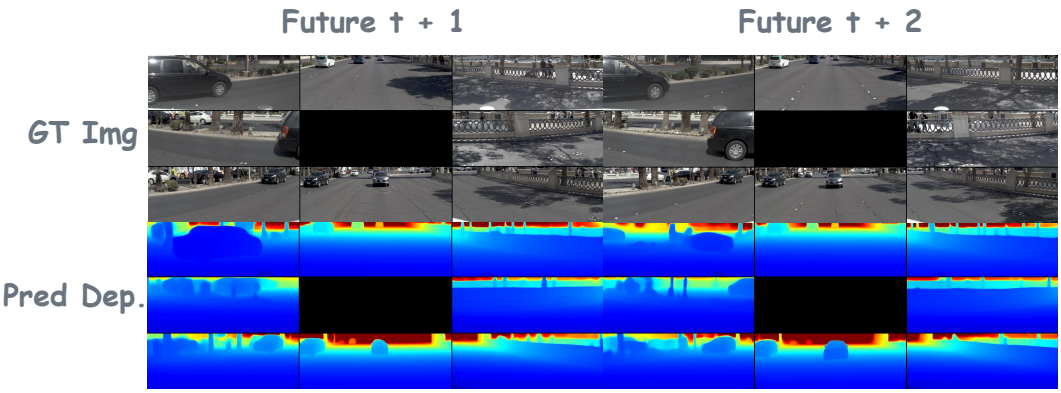


Figure 6: Future depth generation. Ground-truth multi-view images and predicted depth at two future time steps.