2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

ImagiDrive: A Unified Imagination-and-Planning Framework for Autonomous Driving

Jingyu Li[1,2*], Bozhou Zhang[1*], Xin Jin[3], Jiankang Deng[4], Xiatian Zhu[5], Li Zhang[1,2†]

***Abstract*— Autonomous driving requires rich contextual comprehension and precise predictive reasoning to navigate dynamic and complex environments safely. Vision-Language Models (VLMs) and Driving World Models (DWMs) have independently emerged as powerful recipes addressing different aspects of this challenge. VLMs provide interpretability and robust action prediction through their ability to understand multi-modal context, while DWMs excel in generating detailed and plausible future driving scenarios essential for proactive planning. Integrating VLMs with DWMs is an intuitive, promising, yet understudied strategy to exploit the complementary strengths of accurate behavioral prediction and realistic scene generation. Nevertheless, this integration presents notable challenges, particularly in effectively connecting action-level decisions with high-fidelity pixel-level predictions and maintaining computational efficiency. In this paper, we propose *ImagiDrive*, a novel end-to-end autonomous driving framework that integrates a VLM-based driving agent with a DWM-based scene imaginer to form a unified imagination-and-planning loop. The driving agent predicts initial driving trajectories based on multi-modal inputs, guiding the scene imaginer to generate corresponding future scenarios. These imagined scenarios are subsequently utilized to iteratively refine the driving agent's planning decisions. To address efficiency and predictive accuracy challenges inherent in this integration, we introduce an early stopping mechanism and a trajectory selection strategy. Extensive experimental validation on the nuScenes and NAVSIM datasets demonstrates the robustness and superiority of *ImagiDrive* over previous alternatives under both open-loop and closed-loop conditions.**

I. Introduction

End-to-end autonomous driving has made significant progress, with unified models [1], [2], [3], [4], [5] jointly optimizing perception, prediction, and planning on large-scale datasets. Despite strong performance, these methods often lack holistic scene understanding and causal reasoning, limiting their ability to produce rational and flexible trajectories. Recently, vision-language models (VLMs) [6], [7], [8], [9] and driving world models (DWMs) [10], [11], [12], [13], [14], [15] have emerged as promising alternatives. VLMs, pretrained on image-text pairs, offer strong scene comprehension, logical reasoning, and zero-shot generalization, making them ideal for cognitively inspired driving (Fig. 1(a)). Meanwhile, DWMs simulate future scenarios conditioned on past observations and potential actions, enabling agents to anticipate outcomes and evaluate decisions proactively (Fig. 1(b)).

However, integrating these two paradigms remains largely unexplored. DWMs typically focus on improving generative quality in novel scenes, while their potential in action prediction is underutilized. Some methods [14], [16] generate trajectories using inverse dynamics models on generated images, but this process is overly complex and lacks deep scene understanding, often resulting in suboptimal predictions. A natural direction is to combine the complementary strengths of VLMs and DWMs: using planning to guide imagination and imagined futures to refine planning. This integration, however, poses challenges in aligning high-level reasoning with low-level generation and mitigating the slow inference speed of both components.

To this end, we introduce **ImagiDrive**, an end-to-end autonomous driving framework that integrates a VLM-based driving agent with a DWM-based scene imaginer in a recurrent imagination-and-planning loop (Fig. 1(c)). Our driving agent can be easily integrated with mainstream vision-language models such as LLaVA [17], [18] and the InternVL series [7], enabling multi-modal inputs and structured outputs. Meanwhile, our scene imaginer, with its unified input-output design, is capable of generating future scene images conditioned on different driving world models [14], [15]. The core steps of our imagination-and-planning loop are summarized as follows: the agent proposes an initial trajectory from the current frame, which conditions the scene imaginer to generate future scenes. These imagined frames are then fed back into the agent to iteratively refine its planning decisions. To ensure robust and efficient inference, we maintain a trajectory buffer to store trajectories generated in each iteration, and incorporate early stopping and trajectory selection strategies based on safety and consistency.

Our **contributions** are summarized as follows: **(i)** We present *ImagiDrive*, a novel recurrent-loop autonomous driving framework that tightly couples a driving agent with a scene imaginer for imagination-driven planning. **(ii)** We develop a VLM-based driving agent that supports diverse multi-modal inputs and produces structured trajectory predictions. To better integrate with our scene imaginer, we further propose the trajectory buffer with two key strategies: early stopping and trajectory selection, which together enable efficient and reliable inference. **(iii)** Comprehensive experiments on the nuScenes and NAVSIM datasets under both open-loop and closed-loop conditions demonstrate the effectiveness and adaptability of *ImagiDrive*.

† Corresponding author, ∗ Equal contribution
[1] The authors are with School of Data Science, Fudan University.
[2] The authors are with Shanghai Innovation Institute.
[3] The author is with Eastern Institute of Technology.
[4] The author is with Imperial College London.
[5] The author is with University of Surrey.

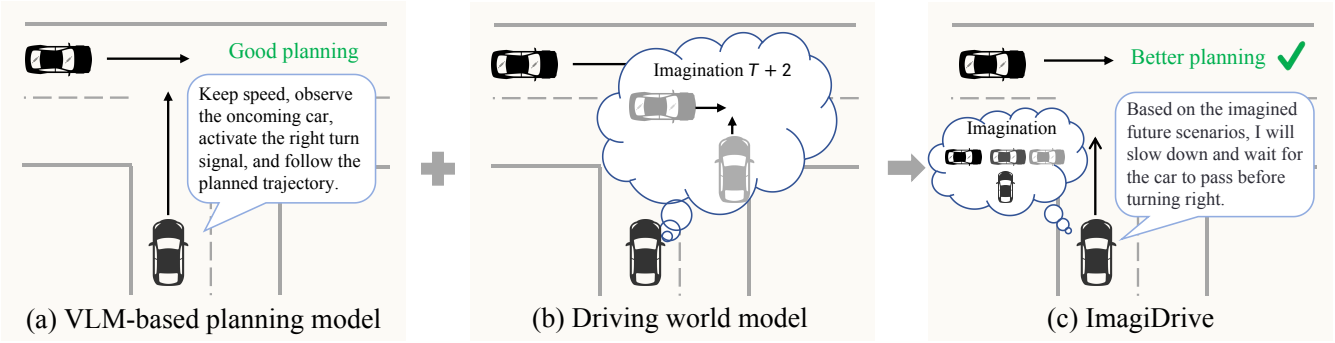

Fig. 1. Overview of autonomous driving paradigms. **(a)** VLM-based end-to-end methods may produce an effective planning strategy to avoid potential collisions. **(b)** DWMs predict and generate future scenarios (T+2 seconds) to identify potential hazards. **(c)** Our proposed framework, ImagiDrive, integrates both paradigms: Using future scene imagination from the DWM-based scene imaginer to iteratively refine VLM-based policy decisions and enhance safety.

II. Related work

VLMs for autonomous driving. Large Language Models (LLMs) [19] and Vision-Language Models (VLMs) [20], [17] exhibit strong multi-modal reasoning capabilities. In autonomous driving, VLMs have emerged as promising human-like agents, benefiting from large-scale datasets with language annotations [21], [22]. Early works [23], [24] leverage general-purpose GPT models but struggle with domain adaptation, prompting the development of domain-specific VLMs for better generalization. DriveLM [25] introduces Chain-of-Thought reasoning from perception to planning, while DriveMM [26] unifies diverse datasets to build a generalizable VLM. Inspired by advances in robotic VLA models [27], recent driving agents like EMMA [6], SimLingo [28], and ORION [29] integrate vision, language, and action for planning and decision-making. However, these methods often rely on complex temporal modeling. In contrast, we propose a VLM that reasons about future behavior via imagined future scenes, without requiring explicit temporal modules.

World models in autonomous driving. World models have become a growing focus in autonomous driving for predicting future scene evolution from current observations. Most approaches adopt generative models, such as autoregressive Transformers (e.g., GAIA-1 [30]) or diffusion models (e.g., DriveDreamer [11], Drive-WM [31]), to synthesize future visual representations. Recent works enhance generation with structured constraints [11], view consistency [32], or large-scale pretraining for zero-shot generalization [16]. However, these models often struggle with physically plausible and detail-rich predictions under complex maneuvers. Vista [14] addresses this by introducing structure-aware losses and larger datasets for improved long-term scene fidelity. Alternatively, some works explore joint generation of future world states and actions [12], [33], or apply world models to end-to-end autonomous driving [34], [35]. In contrast to all these methods, we treat the world model as a future scene synthesizer, conditioned on high-quality trajectories to generate more coherent future scene contexts.

End-to-end autonomous driving. Recent advancements have shifted from isolated task pipelines to unified end-to-end frameworks that jointly address perception and ego-trajectory generation [36], [**?**], [4]. Early works such as ST-P3 [36] adopt intermediate representations to enable end-to-end learning, but limited scene understanding constrains planning quality. UniAD [1] extends this by unifying diverse subtasks, while VAD [2] further improves modularity through vectorized representations. SparseAD [37] and SparseDrive [3] enhance efficiency and scalability via sparse inputs. With more challenging simulators and benchmarks [38], [39], recent efforts have focused on real-world and closed-loop driving. VADv2 [40] achieves state-of-the-art CARLA [41], [42] performance by incorporating a large trajectory vocabulary, conflict-aware loss, and probabilistic planning. DiffusionDrive [5] improves accuracy and diversity with a truncated diffusion policy.

III. Method

Overview. Our *ImagiDrive*, as illustrated in Fig. 2, integrates a driving agent, a trajectory buffer and a scene imaginer in a *recurrent imagination-and-planning* framework: Given a current frame as input, the driving agent predicts an initial trajectory, which is then used to guide the scene imaginer in generating short-term future scene sequences; Subsequently, selected future frames are fed back into the agent for iterative planning refinement. To enhance efficiency and safety, we further introduce a convergence-based early stopping mechanism and a direction-consistent trajectory selection strategy.

A. Driving agent

We build the driving agent based on the VLM with additional input modalities and output heads (Fig. 3). VLM provides a unified and simple framework that can flexibly handle single-frame or multi-frame inputs, incorporate additional driving states, and support interpretable trajectory prediction.

Flexible input integration. The input of our method is multi-modal, consisting of visual data, ego state, textual prompts, and a set of trajectory queries to predict reliable driving plans. Specifically, the visual input can be either

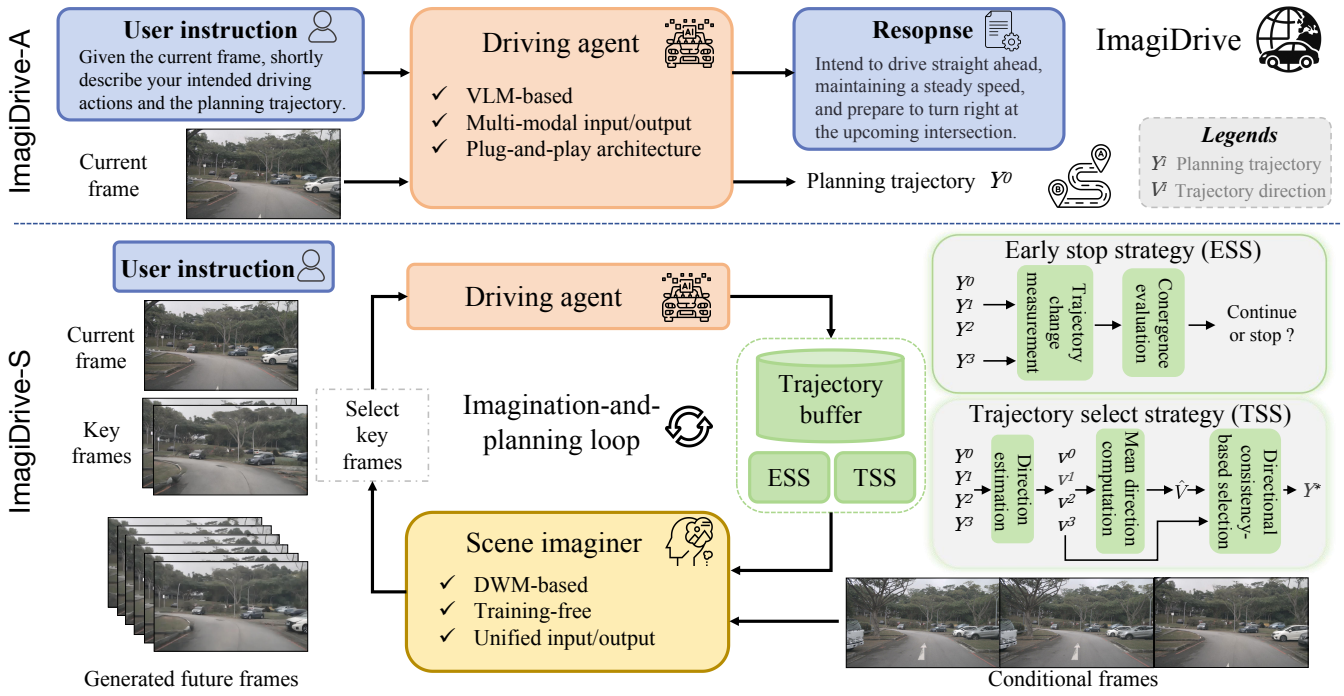

Fig. 2. Overview of *ImagiDrive*. Overview of our system, which includes a driving agent, a scene imaginer, and a trajectory buffer. It operates in two modes: ImagiDrive-A is a standard planning model that uses only the driving agent, while ImagiDrive-S adopts an imagination-and-planning loop, where the scene imaginer generates future frames based on past observations and predicted trajectories. These imagined frames are iteratively fed back to refine planning. The trajectory buffer stores all trajectories, selects the best one, and decides early termination.

a single front-view image or a sequence of frames that includes future predictions generated by the scene imaginer. We encode the vehicle's current speed as an ego state and represent it with a dedicated placeholder token *ego token*, enabling seamless integration into the VLM input stream. Similarly, we allocate several *trajectory token*s as placeholders to indicate positions in the sequence where the model is expected to reason about the trajectory at different future time steps. To enable adaptation to different tasks, such as predicting from the current frame or refining plans with future information, we design two prompt templates. These templates guide effective interaction among visual features, textual prompts, and trajectory tokens for plan generation.

Vision-language model. Thanks to the unified token-based input and output architecture of the current VLMs, our driving agent can seamlessly support both single-frame and multi-frame inputs. For current frame inputs, we feed the image $I_c \in \mathbb{R}^{H\times W}$ directly into the visual encoder. In contrast, future frames $I_f \in \mathbb{R}^{H\times W}$ generated by the scene imaginer frequently exhibit artifacts such as soft edges, ghosting, and diminished texture fidelity. To address this, we apply targeted distortion augmentations during training, including Gaussian blur, shadow overlays, and random noise injection, to improve the model's robustness against such distortions.

Similar to the standard VLM processing pipeline, we use a vision encoder to process image inputs e_i and a tokenizer to handle language instruction inputs e_l. In addition, we employ a lightweight MLP network to process the additional ego state information e_{ego}. We initialize a set of learnable trajectory queries $q \in \mathbb{R}^{N_q\times C}$ to predict trajectory, where each query corresponds to waypoints at different future time steps, and N_q denotes the number of queries. After encoding each modality, we replace the placeholder tokens with their corresponding embeddings to form the final input sequence. The VLM then processes this sequence to produce output features:

$$o_l, \tilde{q} = VLM(e_i, e_{ego}, e_l, q). \tag{1}$$

where o_l denotes the language predictions and $\tilde{q} = \{\tilde{q}_1, \ldots, \tilde{q}_{N_q}\}$ represents the contextualized trajectory queries, which are subsequently sent to the trajectory decoder to get final trajectory.

Unified output format. For the language output, we perform auto-regressive decoding and apply a cross-entropy loss on the predicted tokens. Unlike previous methods [4], [3] that utilize multiple queries to represent diverse candidate trajectories, we design each trajectory query $\tilde{q}_i$ to correspond to a specific future time step t_i. We employ an MLP as the trajectory decoder and the model outputs a sequence of 2D waypoints: $W = \{w_{t_1}, w_{t_2}, \ldots, w_{t_{N_q}}\}$, where each $w_{t_i} \in \mathbb{R}^2$ denotes the predicted vehicle position at time t_i. This design allows the model to generate a temporally coherent trajectory by explicitly reasoning over multiple future steps. We use Smooth L1 loss to supervise the predicted waypoints, and the overall training objective jointly optimizes the language modeling loss $\mathcal{L}_l$ and the trajectory

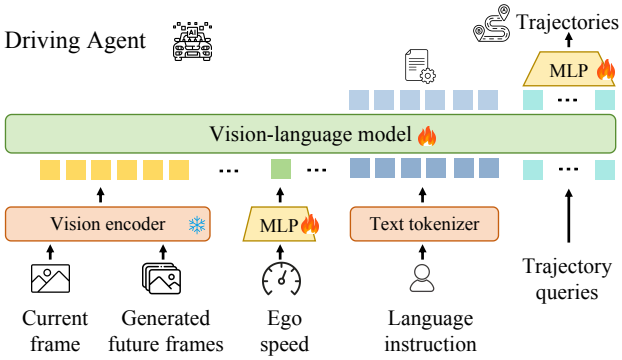

Fig. 3. Overview of *Driving agent*. The agent takes multi-model inputs and produces both language and trajectory predictions.

prediction loss $\mathcal{L}_{\text{traj}}$. The total loss is shown below:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{l}} + \lambda \mathcal{L}_{\text{traj}}, \tag{2}$$

where λ is a weighting factor. This unified decoding framework enables the VLMs to reason holistically over both textual intent and spatiotemporal planning, producing interpretable commands and temporally structured trajectories.

B. Imagination-and-planning loop

Taking advantage of world models [43], [5], [44], [45] in predicting future scene evolution from current observations, we propose to *directly anticipate the future scenes*, bypassing the scene dynamics modeling as done in prior works.

Scene imaginer. Existing generative approaches for autonomous driving can be broadly categorized into diffusion-based [14] and GPT-based methods[33], [46]. Diffusion-based methods focus on generating future frames conditioned on reference images, and rely on future-motion-consistent frames to derive future driving plans via inverse kinematics modeling. In contrast, GPT-based methods decouple image and motion modeling, enabling the joint generation of future visual scenes and driving plans. However, both paradigms face challenges in producing precise and fine-grained future trajectories due to limited capacity for comprehensive scene understanding [15]. Rather than directly addressing these limitations, we leverage the strong generative capabilities of both paradigms through our Scene Imaginer: given conditional images and a continuous trajectory, it generates a temporally coherent sequence of future frames that reflect plausible driving outcomes.

Imagination and planning. Initialization of our method commences by generating a preliminary trajectory $\hat{Y}_{\text{curr}}$, using the VLM-based driving agent conditioned on the current image frame. To enhance the quality and temporal consistency of subsequent predictions by the scene imaginer, its generation process is conditioned on three preceding frames (spanning 0.25 seconds) alongside the current frame. Subsequently, the scene imaginer (SI) utilizes the historical image sequence $\mathbf{I}_{obs}$ and the conditional trajectory $\hat{Y}_{curr}$ to generate a short-term video prediction:

$$\hat{\mathbf{V}}_{fut} = SI(\mathbf{I}_{obs}, \hat{Y}_{curr}) \tag{3}$$

We then select two special frames from the generated future sequence as key frames, corresponding to 0.5 and 1.0 seconds into the future, respectively. These frames, together with the current frame, are used as inputs to the agent to guide subsequent trajectory prediction.

After the initial round of planning and imagination, our model proceeds iteratively: the VLM-based driving agent refines its trajectory by incorporating the current frame and selected future key frames generated by the scene imaginer, shown in Fig. 2. We store each trajectory predicted by the agent into the trajectory buffer. This process continues for a fixed number of iterations or terminates early if a stopping criterion is triggered.

Early stop strategy. As driving agent and scene imaginer both are inefficient, minimizing the computation overhead is crucial for practical deployment. *We observe that when the predicted trajectories converge across consecutive iterations, the corresponding generated future images would become increasingly similar, yielding diminishing returns*. Motivated by this, we design an early stopping metric named *Trajectory Convergence Ratio* (TCR), which enables us to adaptively terminate the iterative process once the trajectory change diminishes. Specifically, TCR captures the normalized rate of change across different time steps of a trajectory, enabling stable and consistent evaluation regardless of the trajectory's scale or magnitude, formulated as:

$$\text{TCR}(\hat{Y}^{(i)}, \hat{Y}^{(j)}) = \frac{1}{N_q} \sum_{t=1}^{N_q} \frac{\left\| \hat{Y}_t^{(i)} - \hat{Y}_t^{(j)} \right\|_2}{\left\| \hat{Y}_t^{(j)} \right\|_2 + \varepsilon}, \tag{4}$$

where $\hat{Y}^{(0)}$ and $\hat{Y}^{(i)}$ denote the initial and the predicted trajectories at iteration i, respectively. N_q is the number of predicted waypoints and ε is a small constant added for numerical stability to avoid division by zero. At each iteration, we compute the TCR between $\hat{Y}^{(i)}$ and all previous predictions $\hat{Y}^{(j)}$ $(j < i)$. The iterative process will terminate once the TCR falls below a predefined threshold θ, indicating sufficient convergence.

Trajectory selection strategy. To obtain more robust trajectories, we design a trajectory selection strategy based on the concept of modal directionality from directional statistics, emphasizing trend direction consistency. Given a set of predicted trajectories $\{\hat{Y}^{(i)}\}_{i=1}^{N}$, we first compute the unit average direction vector for each trajectory:

$$\hat{\mathbf{v}}^{(i)} = \frac{1}{n-1} \sum_{t=1}^{n-1} \frac{\mathbf{p}_{t+1}^{(i)} - \mathbf{p}_t^{(i)}}{\|\mathbf{p}_{t+1}^{(i)} - \mathbf{p}_t^{(i)}\| + \varepsilon}, \tag{5}$$

where p_t is the waypoints at time t. The mean direction vector across all predictions is:

$$\hat{\bar{\mathbf{v}}} = \frac{1}{N} \sum_{i=1}^{N} \hat{\mathbf{v}}^{(i)}, \quad \hat{\bar{\mathbf{v}}} \leftarrow \frac{\hat{\bar{\mathbf{v}}}}{\|\hat{\bar{\mathbf{v}}}\| + \varepsilon}. \tag{6}$$

We then compute the angle between each predicted direction and the global mean. The final selected trajectory is the

TABLE I

CLOSED-LOOP PLANNING RESULTS ON NEURONCAP [47]. † CORRESPONDS TO THE MODEL'S ORIGINAL SETTING. AVG AND STAT REPRESENT AVERAGE AND STATIONARY, RESPECTIVELY. "-": UNREPORTED RESULTS.

Method	NeuroNCAP score ↑				Collision rate (%) ↓			
	Avg.	Stat.	Frontal	Side	Avg.	Stat.	Frontal	Side
UniAD [1]	0.73	0.84	0.10	1.26	88.60	87.80	98.40	79.60
VAD† [2]	0.66	0.47	0.04	1.45	92.50	96.20	99.60	81.60
SparseDrive† [3]	0.92	1.16	0.63	0.96	93.90	93.40	98.40	90.00
BridgeAD [4]	1.60	-	-	-	72.60	-	-	-
Impromptu VLA [48]	2.15	1.77	2.31	1.75	65.50	70.00	**59.00**	65.00
LLava-1.6 + ImagiDrive-A (Ours)	2.37	2.89	1.40	2.81	66.79	50.13	90.06	60.18
LLava-1.6 + ImagiDrive-S (Ours)	2.99	3.63	1.88	3.47	59.45	45.72	82.15	50.49
InternVL2.5 + ImagiDrive-A (Ours)	3.11	3.81	1.97	3.54	48.57	37.24	78.47	30.01
InternVL2.5 + ImagiDrive-S (Ours)	**3.49**	**4.15**	**2.45**	**3.88**	**44.90**	**33.80**	74.00	**26.80**

one *most aligned* with the average direction:

$$\theta^{(i)} = \arccos\left(\text{clip}\left(\hat{\mathbf{v}}^{(i)} \cdot \hat{\bar{\mathbf{v}}}, -1, 1\right)\right), \tag{7}$$

$$i^* = \arg\min_i \theta^{(i)}, \quad \hat{Y}^* = \hat{Y}^{(i^*)}. \tag{8}$$

IV. EXPERIMENTS

A. Datasets

We evaluate our method on three datasets, including both closed-loop and open-loop settings. NeuroNCAP [47] simulator, a photorealistic platform built on nuScenes that enables safety-critical closed-loop evaluation. Turning-nuScenes [49], a smaller yet more challenging dataset. It contains a challenging subset of turning scenes, including 17 scenes with 680 samples. NAVSIM [38] is a large-scale real-world autonomous driving dataset designed for non-reactive simulation and benchmarking. It is built upon OpenScene [50] and focuses on challenging scenarios involving dynamic intention changes, while filtering out trivial cases such as stationary scenes or constant-speed driving.

B. Implementation details

Since our driving agent is easily integrated into VLMs, we select two representative VLM models, LLaVA-1.6-7B [18] and InternVL2.5-4B [7], to serve as our driving agents. Meanwhile, we adopt Vista [14] as the scene imaginer on the nuScenes dataset and Epona [15] as the scene imaginer on the NAVSIM dataset.

For our driving agent, we first train our driving agent with OmniDrive [51] for 3 epochs, facilitating alignment between vision and language in autonomous driving scenarios. We then design a planning QA template with special trajectory tokens and train the model for 5 epochs. We use data augmentation techniques for future frames to simulate the distortion caused by the diffusion model during training. We set the number of trajectory queries as 6, λ_1 and λ_2 as 0.1 and 0.5, respectively. The model is trained on 4 NVIDIA H100 GPUs with a batch size of 1 for 12 hours.

For our Imagination-and-planning loop, we set θ as 0.05 empirically, and the maximum of iterations for the recurrent loop as 5. Our *ImagiDrive* has two variants: **ImagiDrive-A**, a VLM-based agent relying solely on the current frame, and **ImagiDrive-S**, with a scene imaginer to generate and incorporate two future frames, enabling more informed planning.

C. Results analysis

Results on NeuroNCAP. We adopt NeuroNCAP [47] as the closed-loop evaluation benchmark. As shown in Table I, regardless of whether LLaVA-1.6 or InternVL 2.5 is used, our ImagiDrive-A consistently and significantly outperforms the existing state-of-the-art E2E method SparseDrive [3]. Our method significantly outperforms ImpromptuVLA [48], a model trained on extensive additional data, across all metrics with the sole exception of a slightly higher Frontal collision rate. This effectively demonstrates that our agent is capable of understanding the current scene and taking appropriate actions to avoid potential hazards. When using ImagiDrive-S, thanks to our scene imaginer and the imagination-and-planning loop, our method achieves improvements of 0.62 and 0.38 in NeuroNCAP score, along with collision rate reductions of 7.34% and 3.67%, respectively. The improved performance of ImagiDrive-S compared to ImagiDrive-A indicates that incorporating generated future frames as input effectively helps the agent anticipate and avoid potential hazards, fulfilling the goal of integrating imagination and planning in a unified framework.

Results on Turning-nuScenes. We conduct open-loop evaluation on the Turning-nuScenes dataset [49]. Turning-nuScenes is a subset of nuScenes that focuses on turning scenarios. Since the majority of scenes in nuScenes involve straight driving, the turning subset presents greater challenges for planning. A combination of LLaVA-1.6 [18] and ImagiDrive-A slightly underperforms UniAD [1]. With the integration of our scene imaginer, the performance exceeds that of VAD [2], suggesting that incorporating imagined future scenes introduces richer contextual information, leading to more accurate trajectory prediction and notably lower collision rates. Benefiting from the strong performance of InternVL [7], our InternVL2.5-ImagiDrive-A achieves low collision rates and trajectory deviation errors. Moreover, by introducing the imagination-and-planning loop, InternVL2.5-ImagiDrive-S further improves the performance, surpassing MomAD and significantly reducing the collision rate. The results under challenging scenarios further demonstrate that our proposed imagination-and-planning framework effectively leverages the strengths of both the VLM and DWM, leading to more robust and efficient performance.

Results on NAVSIM. We conduct experiments on the

TABLE II

OPEN-LOOP PLANNING RESULTS ON THE TURNING-NUSCENES VALIDATION DATASET [49]. WE FOLLOW THE ST-P3 [36] EVALUATION METRIC.

Method	L2 (m) ↓				Col. Rate (%) ↓			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.
UniAD [1]	0.52	0.88	1.64	1.01	0.16	0.51	1.41	0.69
VAD [2]	0.48	0.80	1.55	0.94	0.07	0.41	1.20	0.56
SparseDrive [3]	0.35	0.77	1.46	0.86	0.04	0.17	0.98	0.40
MomAD [49]	**0.33**	0.70	1.24	0.76	0.03	0.13	0.79	0.32
LLava-1.6 + ImagiDrive-A (Ours)	0.62	1.06	1.71	1.13	0.17	0.41	1.37	0.65
LLava-1.6 + ImagiDrive-S (Ours)	0.48	0.82	1.50	0.93	0.09	0.32	1.08	0.50
InternVL2.5 + ImagiDrive-A (Ours)	0.40	0.81	1.35	0.85	**0.00**	0.28	0.89	0.39
InternVL2.5 + ImagiDrive-S (Ours)	0.34	**0.69**	**1.23**	**0.75**	**0.00**	**0.12**	**0.53**	**0.22**

TABLE III

PLANNING PERFORMANCE ON THE NAVSIM [38] TEST SET EVALUATED WITH CLOSED-LOOP METRICS. NC: NO AT-FAULT COLLISION; DAC: DRIVABLE AREA COMPLIANCE; TTC: TIME-TO-COLLISION; COMF.: COMFORT; EP: EGO PROGRESS; PDMS: PREDICTIVE DRIVER MODEL SCORE. OUR WORLD MODEL OUTPERFORMS STRONG END-TO-END PLANNERS IN TERMS OF OVERALL PDMS. THE UPPER SECTION PRESENTS END-TO-END METHODS TRAINED WITH DENSE SCENE ANNOTATIONS, WHILE THE LOWER SECTION INCLUDES METHODS THAT ARE ASSISTED BY A WORLD MODEL WITHOUT RELYING ON DENSE SUPERVISION.

Method	Input	NC ↑	DAC ↑	TTC ↑	Comf. ↑	EP ↑	PDMS ↑
VADv2 [40]	Camera & Lidar	97.2	89.1	91.9	100	76.0	80.9
TransFuser [52]	Camera & Lidar	97.7	92.8	92.8	100	79.2	84.0
UniAD [1]	Camera	97.8	91.9	92.9	100	78.8	83.4
PARA-Drive [53]	Camera	97.9	92.4	93.0	99.8	79.3	84.6
DRAMA [54]	Camera & Lidar	98.0	93.1	94.8	100	80.1	85.5
DrivingGPT [33]	Camera	**98.9**	90.7	**94.9**	95.6	79.7	82.4
World4Drive [55]	Camera	97.4	94.3	92.8	**100.0**	79.9	85.1
Epona [15]	Camera	97.9	95.1	93.8	99.9	80.4	86.2
LLava-1.6 + ImagiDrive-A (Ours)	Camera	97.7	95.3	93.0	99.9	80.4	86.0
LLava-1.6 + ImagiDrive-S (Ours)	Camera	97.9	95.5	93.1	99.9	**80.7**	86.4
InternVL2.5 + ImagiDrive-A (Ours)	Camera	98.1	**96.2**	94.4	**100.0**	80.1	86.9
InternVL2.5 + ImagiDrive-S (Ours)	Camera	98.6	**96.2**	94.5	**100.0**	80.5	**87.4**

TABLE IV

ABLATIONS ON EARLY STOP STRATEGY (ESS) AND TRAJECTORY SELECTION STRETEGY (TSS).

ESS	TSS	Avg. Col. Rate (%) ↓	Avg. Iterations (steps) ↓
		0.39	5
✓		0.28	2.3
	✓	0.22	5
✓	✓	0.21	2.3

TABLE V

ABLATION STUDY ON DIFFERENT TRAJECTORY SELECTION STRATEGIES.

Method	Col. Rate (%) ↓			
	2.0s	2.5 s	3.0s	*Avg.*
SmoothSel	0.18	0.37	0.58	0.38
SoftMin	0.23	0.44	0.67	0.45
MaxCons	0.18	0.37	0.61	0.39
Ours	**0.14**	**0.33**	**0.55**	**0.34**

NAVSIM [38] dataset with closed-loop metrics, as shown in Table III. We compare our ImagiDrive with some end-to-end methods [40], [1]. Leveraging the VLM's superior scene understanding capabilities, ImagiDrive-A significantly outperforms end-to-end methods on PDMS. Furthermore, ImagiDrive-S fully exploits the generative capability of the scene imaginer and the reasoning strength of the VLM through the imagination-and-planning loop, leading to superior performance on the EP edge cases compared to purely world model-based approaches [15], [55].

D. Ablation study

Main ablation study. We conduct ablation studies to effect of early stop strategy and trajectory selection strategy on the Turning-nuScenes [49]. As shown in Table IV, We report the results in terms of collision rate and iteration steps per trajectory. Without either strategy, the average collision rate is 0.39%, and the average number of iterations per trajectory is 5. While applying ESS alone reduces inference time significantly, it slightly increases collision rate. In contrast, TSS alone improves collision avoidance without affecting runtime. Combining both strategies leads to a favorable trade-off: achieving low collision rate while nearly halving the number of iterations. This validates the effectiveness of our design in balancing safety and efficiency.

Ablation study on Trajectory Selection Strategies. We compare our trajectory selection strategy on Turning-nuScene [49] with other strategies and the results are shown in Table V. 1) Smoothness-Based Selection (MaxCons), this method selects the trajectory with the lowest curvature variation, encouraging smooth and continuous motion patterns. 2) Soft-Min Weighted Averaging (SoftMin), this approach performs a soft-min weighted average over all candidate trajectories based on a predefined cost metric. 3) Trajectory with Maximum Directional Consistency (MaxCons), this strategy selects the trajectory that best aligns with the historical motion direction, favoring smoother transitions. Our method achieves the lowest collision rates across all time horizons, with an average of 0.34%, outperforming all baseline

Fig. 4. Qualitative results in the closed-loop evaluation demonstrate that our ImagiDrive effectively avoids collisions in intersection side-encounter scenario.

Fig. 5. Qualitative results from the open-loop evaluation show that our ImagiDrive can effectively correct the trajectory.

strategies. While SmoothSel and MaxCons offer relatively stable performance, SoftMin lags behind. This highlights the effectiveness of our selection strategy in enhancing safety under long-term prediction.

E. Qualitative analysis

We illustrate the closed-loop results for a safety-critical scenario. As shown in Fig 4, our ImagiDrive is capable of timely deceleration to avoid obstacles upon detecting other vehicles, and gradually accelerates back to normal speed after completing the passing maneuver. We also present qualitative results in open-loop scenarios, as shown in Fig 5, ImagiDrive-A exhibits a significant deviation from the ground-truth trajectory when predicting a right turn, which may lead to a potential collision. In contrast, ImagiDrive-S, with the aid of the world model to generate future visual context, effectively corrects the trajectory.

V. Conclusion

We propose *ImagiDrive*, a novel end-to-end autonomous driving framework that unifies a driving agent and a scene imaginer into an imagination-and-planning paradigm. The system forms a recurrent loop: it predicts trajectories from a single image, generates future images conditioned on the trajectory, and refines the trajectory using these imagined frames. To enhance efficiency, we introduce an early stopping strategy and a trajectory selection strategy. ImagiDrive achieves compelling results in closed-loop evaluations and outperforms previous methods on the challenging Turning-nuScenes and NAVSIM dataset, reflecting its strength in complex scenarios.

VI. Acknowledgements

This work was supported in part by New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0123004), Ningbo grant (2025Z038) and National Natural Science Foundation of China (Grant No. 62376060).

References

[1] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, *et al.*, "Planning-oriented autonomous driving," in *CVPR*, 2023.

[2] B. Jiang, S. Chen, Q. Xu, B. Liao, J. Chen, H. Zhou, Q. Zhang, W. Liu, C. Huang, and X. Wang, "Vad: Vectorized scene representation for efficient autonomous driving," in *ICCV*, 2023.

[3] W. Sun, X. Lin, Y. Shi, C. Zhang, H. Wu, and S. Zheng, "Sparsedrive: End-to-end autonomous driving via sparse scene representation," in *ICRA*, 2025.

[4] B. Zhang, N. Song, X. Jin, and L. Zhang, "Bridging past and future: End-to-end autonomous driving with historical prediction and planning," in *CVPR*, 2025.

[5] B. Liao, S. Chen, H. Yin, B. Jiang, C. Wang, S. Yan, X. Zhang, X. Li, Y. Zhang, Q. Zhang, and X. Wang, "Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving," in *CVPR*, 2025.

[6] J.-J. Hwang, R. Xu, H. Lin, W.-C. Hung, J. Ji, K. Choi, D. Huang, T. He, P. Covington, B. Sapp, *et al.*, "Emma: End-to-end multimodal model for autonomous driving," *arXiv preprint arXiv:2410.23262*, 2024.

[7] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, *et al.*, "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *CVPR*, 2024.

[8] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[9] H. Shao, Y. Hu, L. Wang, G. Song, S. L. Waslander, Y. Liu, and H. Li, "Lmdrive: Closed-loop end-to-end driving with large language models," in *CVPR*, 2024.

[10] J. Lu, Z. Huang, Z. Yang, J. Zhang, and L. Zhang, "Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation," in *ECCV*, 2024.

[11] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu, "Drivedreamer: Towards real-world-drive world models for autonomous driving," in *ECCV*, 2024.

[12] W. Zheng, W. Chen, Y. Huang, B. Zhang, Y. Duan, and J. Lu, "Occworld: Learning a 3d occupancy world model for autonomous driving," in *ECCV*, 2024.

[13] L. Zhang, Y. Xiong, Z. Yang, S. C. ROMERO, and R. Urtasun, "Learning unsupervised world models for autonomous driving via discrete diffusion," in *ICLR*, 2024.

[14] S. Gao, J. Yang, L. Chen, K. Chitta, Y. Qiu, A. Geiger, J. Zhang, and H. Li, "Vista: A generalizable driving world model with high fidelity and versatile controllability," in *NeurIPS*, 2024.

[15] K. Zhang, Z. Tang, X. Hu, X. Pan, X. Guo, Y. Liu, J. Huang, L. Yuan, Q. Zhang, X.-X. Long, *et al.*, "Epona: Autoregressive diffusion world model for autonomous driving," *arXiv preprint arXiv:2506.24113*, 2025.

[16] J. Yang, S. Gao, Y. Qiu, L. Chen, T. Li, B. Dai, K. Chitta, P. Wu, J. Zeng, P. Luo, *et al.*, "Generalized predictive model for autonomous driving," in *CVPR*, 2024.

[17] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *NeurIPS*, 2024.

[18] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, "Llava-next: Improved reasoning, ocr, and world knowledge," 2024. [Online]. Available: https://llava-vl.github.io/blog/2024-01-30-llava-next/

[19] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[20] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[21] M. Nie, R. Peng, C. Wang, X. Cai, J. Han, H. Xu, and L. Zhang, "Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving," in *ECCV*, 2024.

[22] T. Qian, J. Chen, L. Zhuo, Y. Jiao, and Y.-G. Jiang, "Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario," in *AAAI*, 2024.

[23] J. Mao, Y. Qian, J. Ye, H. Zhao, and Y. Wang, "Gpt-driver: Learning to drive with gpt," *arXiv preprint arXiv:2310.01415*, 2023.

[24] L. Wen, D. Fu, X. Li, X. Cai, T. Ma, P. Cai, M. Dou, B. Shi, L. He, and Y. Qiao, "Dilu: A knowledge-driven approach to autonomous driving with large language models," in *ICLR*, 2024.

[25] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, P. Luo, A. Geiger, and H. Li, "Drivelm: Driving with graph visual question answering," in *ECCV*, 2024.

[26] Z. Huang, C. Feng, F. Yan, B. Xiao, Z. Jie, Y. Zhong, X. Liang, and L. Ma, "Drivemm: All-in-one large multimodal model for autonomous driving," *arXiv preprint arXiv:2412.07689*, 2024.

[27] M. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[28] K. Renz, L. Chen, E. Arani, and O. Sinavski, "Simlingo: Vision-only closed-loop autonomous driving with language-action alignment," in *CVPR*, 2025.

[29] H. Fu, D. Zhang, Z. Zhao, J. Cui, D. Liang, C. Zhang, D. Zhang, H. Xie, B. Wang, and X. Bai, "Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation," *arXiv preprint arXiv:2503.19755*, 2025.

[30] A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, and G. Corrado, "Gaia-1: A generative world model for autonomous driving," *arXiv preprint arXiv:2309.17080*, 2023.

[31] Y. Wang, J. He, L. Fan, H. Li, Y. Chen, and Z. Zhang, "Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving," in *CVPR*, 2024.

[32] R. Gao, K. Chen, E. Xie, H. Lanqing, Z. Li, D.-Y. Yeung, and Q. Xu, "Magicdrive: Street view generation with diverse 3d geometry control," in *ICLR*, 2024.

[33] Y. Chen, Y. Wang, and Z. Zhang, "Drivinggpt: Unifying driving world modeling and planning with multi-modal autoregressive transformers," *arXiv preprint*, 2024.

[34] Y. Yang, J. Mei, Y. Ma, S. Du, W. Chen, Y. Qian, Y. Feng, and Y. Liu, "Driving in the occupancy world: Vision-centric 4d occupancy forecasting and planning via world models for autonomous driving," in *AAAI*, 2025.

[35] X. Yang, L. Wen, Y. Ma, J. Mei, X. Li, T. Wei, W. Lei, D. Fu, P. Cai, M. Dou, B. Shi, L. He, Y. Liu, and Y. Qiao, "Drivearena: A closed-loop generative simulation platform for autonomous driving," *arXiv preprint arXiv:2408.00415*, 2024.

[36] S. Hu, L. Chen, P. Wu, H. Li, J. Yan, and D. Tao, "St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning," in *ECCV*, 2022.

[37] D. Zhang, G. Wang, R. Zhu, J. Zhao, X. Chen, S. Zhang, J. Gong, Q. Zhou, W. Zhang, N. Wang, *et al.*, "Sparsead: Sparse query-centric paradigm for efficient end-to-end autonomous driving," *arXiv preprint arXiv:2404.06892*, 2024.

[38] D. Dauner, M. Hallgarten, T. Li, X. Weng, Z. Huang, Z. Yang, H. Li, I. Gilitschenski, B. Ivanovic, M. Pavone, *et al.*, "Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking," in *NeurIPS*, 2024.

[39] X. Jia, Z. Yang, Q. Li, Z. Zhang, and J. Yan, "Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving," in *NeurIPS*, 2024.

[40] S. Chen, B. Jiang, H. Gao, B. Liao, Q. Xu, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Vadv2: End-to-end vectorized autonomous driving via probabilistic planning," *arXiv preprint arXiv:2402.13243*, 2024.

[41] X. Jia, Y. Gao, L. Chen, J. Yan, P. L. Liu, and H. Li, "Driveadapter: Breaking the coupling barrier of perception and planning in end-to-end autonomous driving," in *ICCV*, 2023.

[42] X. Jia, P. Wu, L. Chen, J. Xie, C. He, J. Yan, and H. Li, "Think twice before driving: Towards scalable decoders for end-to-end autonomous driving," in *CVPR*, 2023.

[43] F. Jia, W. Mao, Y. Liu, Y. Zhao, Y. Wen, C. Zhang, X. Zhang, and T. Wang, "Adriver-i: A general world model for autonomous driving," *arXiv preprint arXiv:2311.13549*, 2023.

[44] X. Guo, C. Ding, H. Dou, X. Zhang, W. Tang, and W. Wu, "Infinity-drive: Breaking time limits in driving world models," *arXiv preprint arXiv:2412.01522*, 2024.

[45] X. Li, Y. Zhang, and X. Ye, "Drivingdiffusion: Layout-guided multi-view driving scenarios video generation with latent diffusion model," in *ECCV*, 2025.

[46] X. Hu, W. Yin, M. Jia, J. Deng, X. Guo, Q. Zhang, X. Long, and P. Tan, "Drivingworld: Constructing world model for autonomous driving via video gpt," *arXiv preprint arXiv:2412.19505*, 2024.

[47] W. Ljungbergh, A. Tonderski, J. Johnander, H. Caesar, K. Åström, M. Felsberg, and C. Petersson, "Neuroncap: Photorealistic closed-loop safety testing for autonomous driving," in *ECCV*, 2024.

[48] H. Chi, H. ang Gao, Z. Liu, J. Liu, C. Liu, J. Li, K. Yang, Y. Yu, Z. Wang, W. Li, L. Wang, X. Hu, H. Sun, H. Zhao, and H. Zhao, "Impromptu vla: Open weights and open data for driving vision-language-action models," 2025. [Online]. Available: https://arxiv.org/abs/2505.23757

[49] Z. Song, C. Jia, L. Liu, H. Pan, Y. Zhang, J. Wang, X. Zhang, S. Xu, L. Yang, and Y. Luo, "Don't shake the wheel: Momentum-aware planning in end-to-end autonomous driving," in *CVPR*, 2025.

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

[51] S. Wang, Z. Yu, X. Jiang, S. Lan, M. Shi, N. Chang, J. Kautz, Y. Li, and J. M. Alvarez, "Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning," in *CVPR*, 2025.

[52] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *CVPR*, 2021.

[53] X. Weng, B. Ivanovic, Y. Wang, Y. Wang, and M. Pavone, "Para-drive: Parallelized architecture for real-time autonomous driving," in *CVPR*, 2024.

[54] C. Yuan, Z. Zhang, J. Sun, S. Sun, Z. Huang, C. D. W. Lee, D. Li, Y. Han, A. Wong, K. P. Tee, *et al.*, "Drama: An efficient end-to-end motion planner for autonomous driving with mamba," *arXiv preprint*, 2024.

[55] Y. Zheng, P. Yang, Z. Xing, Q. Zhang, Y. Zheng, Y. Gao, P. Li, T. Zhang, Z. Xia, P. Jia, *et al.*, "World4drive: End-to-end autonomous driving via intention-aware physical latent world model," *arXiv preprint*, 2025.