LMGenDrive: Bridging Multimodal Understanding and Generative World Modeling for End-to-End Driving

Hao Shao[1], Letian Wang[2], Yang Zhou[2], Yuxuan Hu[1], Zhuofan Zong[1], Steven L. Waslander[2], Wei Zhan[3], and Hongsheng Li[1,4,5]

[1] CUHK MMLab, The Chinese University of Hong Kong, Hong Kong SAR
[2] University of Toronto, Canada
[3] UC Berkeley, USA
[4] Shenzhen Loop Area Institute, China
[5] CPII under InnoHK, Hong Kong SAR

Abstract. Recent years have witnessed remarkable progress in autonomous driving, yet generalization to long-tail and open-world scenarios remains the primary bottleneck for large-scale deployment. To address this, one line of research explores LLMs and VLMs for their vision-language understanding and reasoning capabilities, equipping AVs with the ability to interpret rare and safety-critical situations when generating driving actions. In parallel, another line investigates generative world models to capture the spatio-temporal evolution of driving scenes, enabling agents to imagine and evaluate possible futures before acting. Inspired by human intelligence, which seamlessly unites understanding and imagination as a hallmark of AGI, this work explores a unified model that brings these two capabilities together for autonomous driving. We present LMGenDrive, the first framework that unifies LLM-based multimodal understanding with generative world models for end-to-end closed-loop autonomous driving. Given multi-view camera inputs and natural-language instructions, our model generates both realistic future driving videos and corresponding control signals. By coupling an LLM with generative video capabilities, LMGenDrive gains complementary benefits: future video prediction enhances spatio-temporal scene modeling, while the LLM provides strong semantic priors and instruction grounding learned from large-scale pretraining. A progressive three-stage training strategy—ranging from vision pretraining to multi-step long-horizon driving—is proposed to further improve stability and performance. The resulting model can also operate in two complementary modes: low-latency online planning and autoregressive offline video generation. Experiments show that LMGenDrive significantly outperforms state-of-the-art methods on challenging closed-loop driving benchmarks, improving instruction following, spatio-temporal understanding, and robustness to rare scenarios. Our work not only sets a new state-of-the-art in autonomous driving, but also demonstrates that unifying multimodal understanding and generation provides a promising direction for building more generalizable and robust embodied decision-making systems.

Keywords: End-to-End Autonomous Driving · World Action Models · Large Language Models

1 Introduction

Remarkable progress in autonomous driving has been witnessed in recent years with an increasing number of commercial autonomous vehicles (AVs) deployed on public roads. Amidst this momentum, end-to-end autonomous driving has emerged as a particularly vibrant research direction. Unlike traditional modular pipelines that separately handle perception, prediction, and planning with handcrafted interfaces, end-to-end models provide a holistic paradigm with the potential to remove information bottlenecks among modules, better align model optimization with system-level performance, and scale effectively with large amounts of driving data.

Despite this progress, the problem of generalization remains the central bottleneck for the entire autonomous driving community. As we approach the frontier of real-world deployment, the ability to robustly handle long-tail edge cases and operate in open-world settings remains the defining challenge for AV systems. These scenarios can include rare but safety-critical events, distribution shifts across regions, adversarial weather conditions, as well as complex social interactions and ambiguous intent among agents. This challenge manifests across the autonomy stack: perception systems struggle to identify open-set entities, while prediction and planning models falter in extrapolating to nondeterministic and previously unseen behaviors. These generalization failures represent the last barrier between research prototypes and truly scalable, globally deployable autonomous vehicles.

Amid this backdrop, large language models (LLMs) have demonstrated strong capabilities in language understanding and structured decision modeling across a wide range of tasks. Recent models such as GPT-5 and DeepSeek-R1 [15] showcase robust capabilities in commonsense reasoning, abstraction, and decision-making. Meanwhile, vision-language models (VLMs) further extend this capacity to the multimodal domain, enabling unified interpretation of textual and visual inputs [1, 2, 27, 31, 51]. Inspired by these vision-language understanding and reasoning capabilities, a wave of research has begun exploring how to equip AV systems with LLMs and VLMs to address the open-world, long-tail challenges in autonomous driving. As exemplified in works such as LMDrive [41] and GPT-Driver [33], these models act as the cognitive brains to interpret ambiguous scenarios and guide complex decision-making under uncertainty. However, most existing LLM- or VLM-empowered driving methods follow the paradigm that maps inputs directly to actions, falling short in explaining and capturing the temporal evolution of driving scenes—an essential factor for robust and anticipatory planning.

Meanwhile, another stream of research, world modeling [17], has emerged to simulate the spatio-temporal evolution of the scenes, as exemplified by video-based works such as Genie-3 [3] and Pandora [57]. Their potential has also been

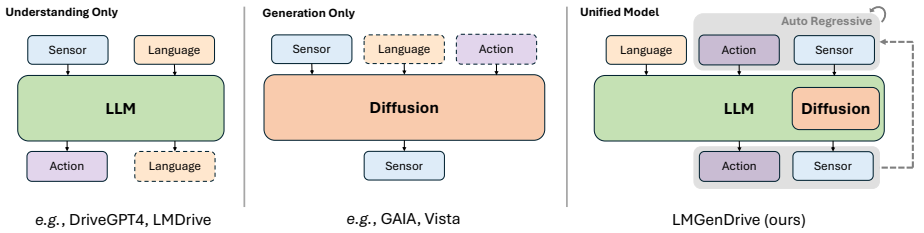

Fig. 1: Comparison between existing works and ours. Prior works either leverage LLMs/VLMs for multimodal understanding and reasoning, or adopt world models for video-based scene imagination, but treat these capabilities in isolation. In contrast, our proposed **LMGenDrive** unifies both within a closed-loop end-to-end framework: the LLM interprets and reasons over multimodal inputs, while the world model simulates future scene evolution, together enabling instruction-guided planning, spatio-temporal understanding, and robust long-horizon driving.

actively explored in the autonomous driving domain, enabling the agent to "imagin" different futures before committing to a plan. However, existing works either focus on solely generating high-fidelity scenes [12, 13, 20, 23, 39, 42, 53, 60, 67], or utilize world models as a plug-and-play forecasting module to rank multiple possible plans [52,54]. The integration of joint video generation and motion planning remains underexplored, limiting their ability to address critical challenges such as cumulative control errors, human-robot interaction, and the temporal consistency between generated actions and videos—factors that are essential for long-horizon problem solving in real-world systems. Moreover, these models generally lack the rich knowledge priors and instruction-following capabilities provided by LLMs.

In contrast to existing models that specialize in either understanding or generation, human cognition naturally integrates perception and imagination to support long-horizon decision-making. Inspired by this observation, we explore whether combining multimodal understanding with generative world modeling can improve robustness and temporal consistency in autonomous driving systems. While recent studies [6, 9, 29, 45] have shown encouraging results combining multimodal understanding and generation within a single model, whether this principle extends to embodied agents—and autonomous driving in particular—remains an open challenge. In this work, we propose LMGenDrive, the first framework that unifies LLM-based multimodal understanding with generative world models for closed-loop end-to-end autonomous driving. Our unified model takes multi-view camera data and natural-language driving instructions as inputs, generating both multi-view future driving videos and control signals for the following timesteps. Concretely, we integrate an LLM and a diffusion-based video generation model: the LLM interprets and fuses visual observations with language instructions, producing learnable queries that capture the evolving scene states, which then serve as conditioning signals for the diffusion model to generate realistic multi-view driving futures. Within this unified architec-

ture, video generation enhances the LLM's spatio-temporal understanding, while the LLM provides instruction-following and reasoning capabilities to the world model, together enabling more robust closed-loop driving.

To enable such a unified model, we also propose a three-stage curriculum training strategy for enhanced performance and stability. First, we pretrain a vision encoder for robust driving scene understanding. Next, the frozen encoder is integrated with the LLM and video generator, and fine-tuned on single-step prediction to ground instruction following and immediate action outcomes. Finally, training is extended to multi-step sequences, enhancing long-horizon temporal modeling and consistency for continuous driving scenarios. Once trained, the model can be applied in two modes: (1) Online planning mode: the model solely predicts planning outputs, with the diffusion generation component discarded to reduce latency; (2) Offline data generation mode: the model conducts autoregressive video generation, where the generated video and predicted control signal serve as input for the next timestep, enabling extended and consistent driving video sequences.

To sum up, our contributions are threefold:

- **Unified closed-loop framework.** We present LMGenDrive, the first framework that unifies LLM-based multimodal understanding with generative world models for closed-loop end-to-end autonomous driving, bridging perception and imagination within a single architecture.
- **Progressive training and dual modes.** We introduce a three-stage training pipeline, from vision pretraining to long-horizon multi-step driving, and support two usage modes: low-latency online planning and offline autoregressive video generation.
- **State-of-the-art performance and empirical insights.** LMGenDrive achieves state-of-the-art closed-loop performance on challenging autonomous driving benchmarks, improving instruction following, spatio-temporal reasoning, and robustness to long-tail scenarios. These results highlight the complementary benefits of combining multimodal understanding with generative world modeling.

2 Related works

2.1 End-to-end driving

Much progress has been made in end-to-end autonomous driving, with many recent methods based on imitation learning. UniAD [21] unified full-stack driving tasks through query-based interfaces, while ThinkTwice [25] retrieved critical-region information to refine predictions [66]. InterFuser [43] used transformers to fuse multi-modal, multi-view sensor data for richer scene understanding. ReasonNet [44] leveraged both temporal and global information of the driving scene to enhance perception, particularly in occlusion scenarios. Para-Drive [55] proposed a fully parallel architecture with a shared BEV representation, and DriveTransformer [26] went further by discarding BEV features and using pure transformers to aggregate sensor and query information. Diffusion models have also been

explored for modeling diverse driving behaviors. DiffusionPlanner [65] learns a diffusion-based driving policy, while DiffAD [50] formulates perception and decision-making as conditional image generation. More recent frameworks, such as SimLingo [36] and Orion [11], incorporate large language models or generative world modeling into the driving loop, enabling richer semantic grounding and action-conditioned scene reasoning. Despite these advances, most approaches still struggle with rare corner cases and lack the reasoning ability needed to generalize beyond the training distribution.

2.2 MLLM for autonomous driving

Recent advances in large language models (LLMs) [15, 22, 24, 48, 49, 59] and vision–language models (VLMs) [2,30,31,35,42,51,68,69] have motivated integrating MLLMs into autonomous driving for stronger reasoning and explainability. Early works like GPT-Driver [33] and LanguageMPC [40] convert driving scenes into textual inputs for direct reasoning. LLM-Driver [7] utilized the numeric vector as the input modality and fused vectorized object-level 2D scene representation to enable the LLM to answer the questions based on the current environment. Later methods employ VLMs to process images and videos: some focus on visual question answering for scene understanding and optional action output (*e.g.*, DriveLM [46], DriveGPT4 [58], DriveVLM [47]), while others predict driving actions end-to-end (*e.g.*, LMDrive [41], DriveMoE [61], BEVDriver [56]). Agentic designs with hierarchical control, tool use, and memory, such as Agent-Driver [34] and AD-H [63], further extend capabilities. However, most MLLM-based approaches emphasize planning or explanation and lack robust modeling of how scenes and surrounding objects evolve over time—a key requirement for anticipating events and ensuring safe, long-horizon decision-making.

2.3 World models for autonomous driving

The concept of a *world model*, a predictive model that simulates environment dynamics, has regained attention. Video generation has become a leading paradigm, supported by advances in generative modeling, large-scale video datasets, and with wide applicability. In autonomous driving, temporally grounded video prediction provides rich context for understanding and decision-making. Several methods treat pure video generation as world modeling. GAIA [20] conditions generation on image, text, and action inputs. GAIA-2 [39] extends this to multi-view scenes, and MagicDrive [12] adds control signals such as HD maps and bounding boxes. Vista [13] scales to internet-scale driving data, while CoGen3D [23] predicts 3D-consistent representations before video synthesis to improve spatial coherence. Beyond pure generation, DriveWM [54] predicts alternative futures for conditional planning. More recent work—DriveDreamer [53], GenAD [60], and ProphetDWM [52]—jointly models videos and actions but still mainly uses open-loop settings without direct feedback. The most related work is LAW [28], which combines world modeling with closed-loop planning. However, it supervises the world model only with next-frame hidden features rather than full

video generation, limiting its ability to simulate realistic scenes or generate synthetic data. Moreover, it lacks language model integration and therefore cannot support instruction following or natural language grounding in driving. To our knowledge, this is the first framework to unify LLM-based reasoning with video world models for closed-loop end-to-end autonomous driving.

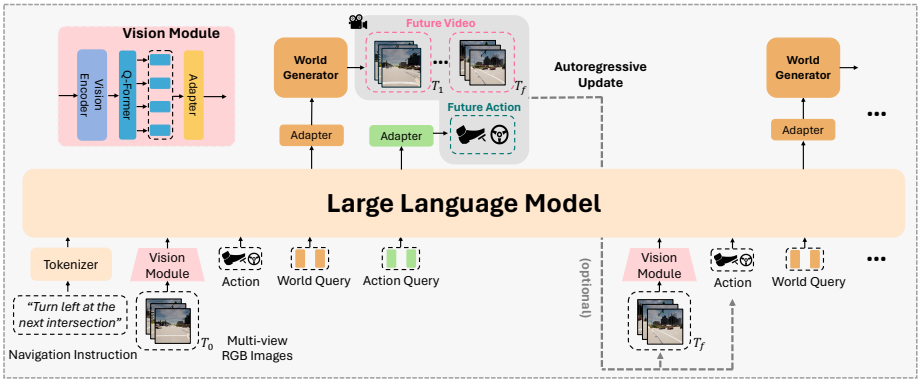

Fig. 2: Overview of our unified understanding and generation architecture. We start by encoding the language instruction, multi-view RGB images, and the current action into the LLM. Two sets of learnable queries, world query and action query, are then fed into the LLM, and ultimately used to generate the future driving video and corresponding actions. The framework supports two operation modes: (1) offline data generation mode, an autoregressive generation process is adopted, where the last frame of the future video and the predicted action are used as inputs for the next timestep; (2) online planning mode, real-world data are provided as inputs for the following timestep.

3 Method

3.1 Overall Framework

In this work, we propose LMGenDrive, a framework that unifies textual understanding/reasoning, future scene generation, and end-to-end planning. As illustrated in Figure 2, LMGenDrive is composed of three major components: (1) a vision encoder that processes multi-view camera sensor data for scene understanding and generating visual tokens; (2) a large language model and its associated component (tokenizer, Q-Former, and adapters) that takes in the language instruction, input visual tokens, world queries, and action queries, to predict the future driving scenes and actions; (3) a multi-view world generator that takes future scene tokens from the LLM and multi-view images from the last frame as inputs, to generate future multi-view driving videos. We will introduce the vision encoder in Section 3.2, the LLM with its associated components

in Section 3.3, and the multi-view world generator in Section 3.4. Finally, we describe the training recipe in Section 3.5.

3.2 Vision encoder

The vision encoder is designed to perceive the environment by processing, fusing, and transforming sensor data into visual tokens that can be consumed by the language model. Prior works [22, 41] typically leverage both multi-view images and LiDAR sensor inputs, where the LiDAR inputs are encoded into bird's-eye view (BEV) queries to extract information from multi-view images. However, our setting focuses on autoregressive video generation—where LiDAR is only available at the current frame but not in future frames. As a result, we replace LiDAR inputs with BEV positional encodings, enabling effective perception while maintaining compatibility with future video generation. The vision encoder consists of three parts: (1) In the sensor encoding part, for each image input, a 2D backbone ResNet [18] is applied to extract the image feature map, which is flattened to one-dimensional tokens. Tokens from different views are then fused by a transformer encoder. (2) In the BEV decoder, BEV position encodings serve as $H \times W$ queries to attend to the multi-view image features and generate BEV tokens. In addition, the learnable queries and one extra query generate corresponding waypoint tokens and one traffic-light token, respectively. The three types of visual tokens (BEV, waypoint, and traffic light) will be presented to the LLM, providing rich scene information. (3) Lastly, as the first-stage training, the vision encoder is pretrained on perception tasks (BEV object detection, traffic light recognition, waypoint prediction) by feeding the three types of tokens to additional prediction heads. Three loss terms, including the detection loss [44], the l_1 waypoint loss and the cross-entropy traffic light prediction loss, are applied respectively. Note that, following LMDrive [41], once pretrained, these prediction heads are discarded and the encoder is frozen, serving as the vision encoder for the large language model.

3.3 LLM for instruction-following driving and scene understanding

As depicted in Figure 2, our system casts the LLM as the "brain" of the entire driving pipeline: it ingests sensor tokens emitted by the frozen vision encoder at every frame and parses natural-language commands, to forecast upcoming maneuvers and emits conditioning features for subsequent video generation. We adopt LLaMA [48] as the linguistic architecture due to its broad success in both language-centric [14,64] and vision-grounded [31,68] instruction-tuning settings.
Instruction and visual tokenization. As the model takes navigation instruction and multi-view image as inputs, their tokenization is our first step. For the navigation instruction, we tokenize them with the LLaMA tokenizer [48]. For the multi-view images, each frame is tokenized by the aforementioned vision encoder, and the resulting tokens are buffered together with the most recent token history (up to $T_{\max}$ frames) to curb cumulative error and maintain temporal coherence during executing the driving instruction in the closed loop. For each frame, the

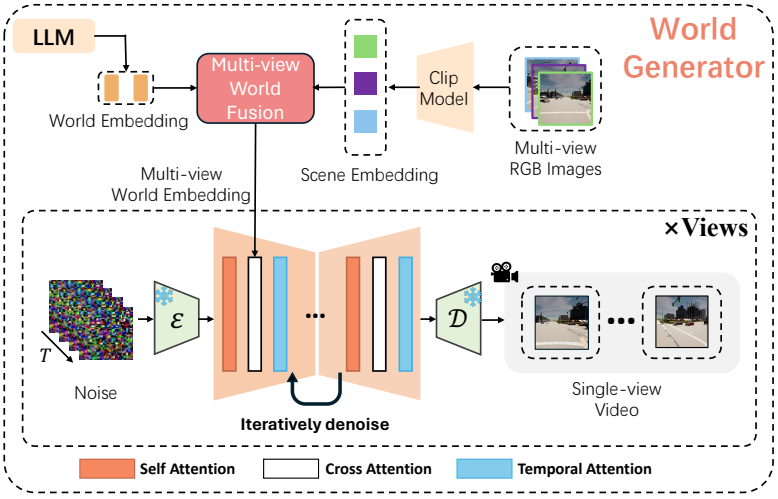

Fig. 3: Architecture of our world generator. We begin by fusing the world embedding obtained from the LLM with multi-view RGB images. The fused multi-view world embedding is then injected into the diffusion model with the cross-attention mechanism to generate multi-view future videos.

pretrained vision encoder outputs $H \times W$ BEV tokens, 4 waypoint tokens, and 1 traffic-light token. Passing all visual tokens (about 2k per frame) to the LLM is computationally prohibitive. To compress them, we use a Q-Former with 8 learnable queries per frame that attend to the raw tokens and distill them into compact frame-level features. An MLP adapter then projects these features to the LLM's embedding dimension for seamless fusion with language tokens.

Action prediction. Together with the instruction and visual tokens, we feed N learnable action query tokens into the LLM. After passing through the LLM and associated adapters, these queries evolve into N latent feature vectors, each encoding the spatio-temporal context needed for motion planning. A subsequent two-layer MLP maps these N feature vectors to N predicted waypoints and outputs a binary flag indicating whether the current instruction has been completed. Finally, the predicted waypoints are converted into low-level control commands—brake, throttle, and steering—through two independent PID controllers [5] that separately regulate longitudinal velocity and lateral heading, ensuring accurate trajectory tracking.

3.4 Multi-View World Model

While the LLM is responsible for instruction following and reasoning, autonomous driving also requires modeling the visual dynamics of the environment. To this end, we introduce a multi-view world model that generates future video frames conditioned on the LLM outputs. By aligning action predictions with video generation, our framework jointly reasons about both the agent's behavior and the

evolution of the surrounding world. Our video generation process additionally *supports* an autoregressive mode [57], which can be optionally enabled during inference: the future action and frame predicted at the previous timestep can be fed back into the LLM as input for the next prediction.

World Query Conditioning. As shown in Figure 2, in addition to the instructional tokens, visual tokens, and action query mentioned above, the LLM also takes the world query as input. Passing through the LLM, these world queries aggregate information from instructions, sensor inputs, and the actions, thereby enabling the model to form an internal representation of the world dynamics. Conceptually, these queries serve as a bridge to the world's temporal evolution, and act as the conditioning signal fed into the following world generator to synthesize future driving videos.

Multi-view Image Conditioning. As shown in Figure 3, besides using scene queries to capture world evolution, we incorporate the last-frame multi-view images to supply fine-grained appearance details and the initial world state. These images are encoded by a CLIP model into semantically rich features that emphasize visual textures and appearance. During end-to-end training, these CLIP features are fused with LLM representations through attention blocks. Self-attention aggregates and aligns multi-view information into a unified space, and cross-attention injects LLM guidance. This design not only provides an appearance prior for consistent video generation but also encourages the LLM to focus on dynamic, motion-related representations.

World Generator. After the multi-view world fusion step, we obtain a set of multi-view world embeddings, each corresponding to one camera view. Taking these embeddings as the final conditioning feature, our world generator employs a U-Net [38] diffusion architecture to produce future frames. For each view, the associated embedding is injected into the diffusion process through a dedicated cross-attention module, ensuring that view-specific information is effectively transferred. The model follows the standard denoising diffusion process [19]: starting from pure Gaussian noise and progressively removing noise to generate a future video sequence. The output is a video tensor of shape $\mathbb{R}^{v\times t\times h\times w\times 3}$, where v is the number of views, t is the temporal length, and h, w are the spatial resolution. Inspired by [16, 54], we further augment the U-Net blocks with spatio-temporal transformers to better capture temporal dynamics and spatial structure in driving scenes.

3.5 Training Recipe

We adopt a three-stage training strategy to progressively build the model's perception, reasoning, and generation capabilities. This curriculum ensures stable convergence and enables effective long-horizon temporal modeling.

Stage 1: Vision Encoder Pretraining. We first pretrain the vision encoder on single-frame perception tasks using 3M expert-collected frames generated in the CARLA simulator [43]. Perception heads are attached for object detection, traffic light classification, and waypoint regression. After convergence, only the vision encoder is retained and frozen in later stages.

Stage 2: Single-Step Planning and Generation. Next, we jointly fine-tune the LLM and the video generator for single-step prediction. The vision encoder is frozen to reduce memory usage. The model takes as input a single-frame multi-view image, natural language instruction, action queries, and world queries, to predict the next waypoint, the instruction completion flag, and the future driving video. This stage enables the LLM to learn grounded instruction-following and understand how the world evolves under given actions. Simultaneously, the world generator learns to synthesize multi-view driving videos conditioned on the last frame and LLM-generated features.

Stage 3: Multi-Step Long-Horizon Training. We progressively expand training to 2–3-step sequences to strengthen long-horizon reasoning. Specifically, previously generated videos are autoregressively fed as input for the next step's generation. To save memory, the video generator is frozen while gradients still propagate, and the LLM remains fully trainable. This design encourages the LLM to capture temporal dependencies—such as other agents' intentions, speed, and interactions—over extended observation windows. As a result, the LLM develops stronger temporal abstraction and inductive reasoning abilities for dynamic driving scenes.

Training Objectives. We apply three loss terms in the last two stages: (1) l_1 waypoint regression loss; (2) binary classification loss for instruction completion; (3) diffusion loss for video generation:

$$\mathcal{L}_{\mathrm{DM}} = \mathbb{E}_{t,\epsilon}\left[\left\|\epsilon_\theta(\mathbf{z}_t, \mathbf{c}, t) - \epsilon\right\|^2\right],$$

where $\mathbf{z}_t$ is the noisy latent at timestep t, ϵ is the added Gaussian noise, and $\mathbf{c}$ denotes conditioning features from the multi-view image and scene queries.

4 Experiments

4.1 Experiment Setup

Training Details. During training, three synchronized RGB cameras (left, front, right) are resized to 224^2 pixels and sampled at 10 Hz, and an 8-frame temporal window is considered. The network is tasked with predicting four future waypoints at $t + \{0.2, 0.4, 0.6, 0.8\}$ s, along with eight future video frames from $t + 0.1$ s to $t + 0.9$ s in 0.1-second increments. We optimize the model using AdamW optimizer [32] with an initial learning rate of 1×10^{-5} on eight NVIDIA H800 GPUs under DeepSpeed ZeRO-2; convergence is reached in roughly two days. Due to GPU-memory constraints, the third training stage operates on one to three timesteps. We allocate a total of 50k optimization iterations across the three stages, using 20k/20k/10k iterations for Stage-1/Stage-2/Stage-3, respectively. The system uses Vicuna-7B [8] as the LLM backbone, Stable Diffusion 1.5 [37] for image generation, and AnimateDiff [16] for temporal modeling.

Benchmark. We implement and evaluate our approach in the open-source CARLA simulator (version 0.9.10.1) [10] on the LangAuto benchmark [41]. The

Table 1: Performance comparison on the LangAuto benchmark. We report the metrics for 3 evaluation runs. AD-H† leverages an extra model OPT-350M [62] for low-level control.

Methods	LangAuto			LangAuto-Short			LangAuto-Tiny		
	DS ↑	RC ↑	IS ↑	DS ↑	RC ↑	IS ↑	DS ↑	RC ↑	IS ↑
LMDrive [41]	10.7±3.8	16.2±4.9	0.63±0.04	14.2±4.4	20.1±4.4	0.72±0.04	20.1±4.1	24.7±5.1	0.75±0.03
AD-H† [63]	44.0	53.2	0.83	56.1	68.0	0.78	77.5	85.1	0.91
BEVDriver [56]	48.9	59.7	0.82	66.7	77.8	0.87	70.2	81.3	0.87
Ours	**62.2±3.3**	**74.5±4.1**	**0.85±0.04**	**77.1±4.1**	**87.9±3.5**	**0.88±0.03**	**84.1±3.6**	**92.5±4.0**	**0.92±0.04**

LangAuto benchmark comprises test routes spanning eight CARLA towns, diverse weather conditions, and deliberately misleading linguistic cues. It includes three tracks—LangAuto, LangAuto-Short, and LangAuto-Tiny—which differ primarily in route length. During evaluation, each method acts as an agent that controls the vehicle using only natural-language commands and visual observations, and directly outputs the corresponding control signals to interact with the real-time environment. Consistent with LMDrive [41], we report results on each track individually.

Metric. Following the CARLA Leaderboard [4] and LangAuto [41], we report route completion (RC), infraction score (IS), and driving score (DS). RC measures the fraction of the planned route completed before exceeding the deviation tolerance. IS penalizes collisions and traffic-rule violations via a decaying factor. DS, the product of RC and IS, serves as the primary overall indicator of safe and efficient driving. For generated videos, we further evaluate perceptual quality using Fréchet Video Distance (FVD) and Fréchet Inception Distance (FID), which assess temporal consistency and visual realism, respectively.

4.2 SOTA Comparison

The experimental results in Table 1 demonstrate that our method significantly outperforms existing state-of-the-art approaches on the LangAuto benchmark. During testing, we append the most recent sensor history (up to $T_{\max}$ frames) at every timestep to curb cumulative error and maintain temporal coherence. Specifically, LMDrive [41] achieves a driving score (DS) of 10.7 in the LangAuto track, while AD-H [63] and BEVDriver [56] demonstrate improved DS values of 44.0 and 48.9, respectively. Our method further pushes the performance to a higher level, with a DS of 62.2. In terms of route completion (RC) and infraction score (IS), our approach also achieves superior performance across all three tracks: LangAuto, LangAuto-Short, and LangAuto-Tiny, showing the effectiveness of our method in handling more complex driving scenarios with language instructions.

4.3 Ablation studies

Ablation Study on Module Design. As shown in Table 2, we conduct four ablation experiments to quantify the contribution of each key component in

Table 2: Ablation study on the module design for planning performance.

Module design	DS ↑	RC ↑	IS ↑
baseline	**62.2±3.3**	**74.5±4.1**	**0.85±0.04**
w/o world generator	53.4±2.2	65.8±4.2	0.80±0.01
w/o action queries	58.7±3.1	70.4±3.7	0.84±0.02
w/o visual pre-training	54.9±4.5	67.1±4.5	0.81±0.02
w/o stage-3 training	55.6±4.5	68.9±4.5	0.80±0.02

Table 3: Ablation study on the module design for generation performance.

Module design	FID↓	FVD↓
baseline	**6.3**	**286**
w/o multi-view fusion	7.8	371
world queries: 64 → 32	10.1	318
world queries: 64 → 16	11.6	424

our proposed LMGenDrive. (1) **w/o world generator**: Removing the world generator together with its world query sharply degrades DS to 53.4, demonstrating that the multi-view world generator is crucial for enriching the LLM's understanding of spatio-temporal dynamics and strengthening future-scene reasoning. With a stronger understanding of world semantics, the driving agent exhibits markedly improved instruction-following capability and safer decision-making behavior. (2) **w/o action queries**: Replacing learnable action queries with an LMDrive-style autoregressive action prediction lowers DS to 58.7, indicating that explicit action queries provide more structured supervision and lead to more reliable planning. (3) **w/o visual pre-training**: Without the first-stage driving-oriented visual pre-training, DS drops to 54.9, highlighting the importance of injecting driving-specific semantics into the vision encoder to enhance downstream scene understanding. (4) **w/o stage-3 training**: Skipping the Multi-Step Long-Horizon training stage reduces DS to 55.6, confirming that long-horizon temporal modeling is essential for robust reasoning over extended driving contexts. Overall, all ablations show degraded DS, with missing world generator or long-horizon training causing the largest drops, confirming these modules—along with visual pre-training and action queries—are vital for accurate, safe planning in LMGenDrive.

Ablation Study on Generation Module Design. As shown in Table 3, we further investigate how key components of the generation module influence video quality, measured by FID and FVD. (1) **w/o multi-view fusion**: Removing the cross-view fusion increases FID from 6.3 to 7.8 and FVD from 286 to 371, indicating that, without interaction among different camera views, the model struggles to maintain spatial consistency. (2) **world queries choices**: Reducing the number of world queries from 64 to 32 or 16 leads to a clear performance drop, with FID/FVD rising to 10.1/318 and 11.6/424, respectively. Overall, these results demonstrate that multi-view fusion and sufficient world queries are critical for generating coherent, high-quality videos.

Ablation Study on Training Stages. To further analyze the contribution of the multi-stage training strategy, we conduct a controlled ablation where the total number of optimization iterations is fixed to 30k across different training configurations. Specifically, we compare three settings: (1) the full training pipeline, (2) removing Stage-2 and training Stage-3 for 30k iterations, and (3) removing Stage-3 and training Stage-2 for 30k iterations. As shown in Table 4, remov-

Table 4: Training-stage ablation under controlled iterations. We report DS/RC/IS on LangAuto-Tiny and LangAuto-Short.

Setting	LangAuto-Tiny			LangAuto-Short		
	DS ↑	RC ↑	IS ↑	DS ↑	RC ↑	IS ↑
baseline (Stage-2: 2w iter, Stage-3: 1w iter)	**84.1**	**92.5**	**0.92**	**77.1**	**87.9**	**0.88**
w/o Stage-2 (Stage-3: 3w iter)	78.0	87.6	0.90	67.9	81.0	0.85
w/o Stage-3 (Stage-2: 3w iter)	80.1	88.5	0.91	72.3	83.7	0.86

Table 5: Scaling of long-horizon video generation. FID/FVD degrade with longer prediction horizons due to accumulated errors.

Horizon (frames)	8	16	24	32	64	128
FID ↓	6.3	7.4	7.8	9.2	12.8	15.1
FVD ↓	286	340	371	435	610	981

ing Stage-3 results in a larger performance drop on LangAuto-Short compared to LangAuto-Tiny, suggesting that multi-step training is particularly beneficial for longer driving routes. In contrast, Stage-2 mainly improves the stability of short-horizon grounding and action prediction.

4.4 Long-Horizon Video Generation Scaling

To evaluate the scalability of our diffusion-based world model to longer horizons, we assess generation quality under increasing prediction lengths. Specifically, we autoregressively generate multi-view future videos with horizons of $\{8, 16, 24, 32, 64, 128\}$ frames and measure Fréchet Inception Distance (FID) and Fréchet Video Distance (FVD) against ground-truth videos on the same evaluation split, where lower values indicate better quality.

Compared with existing driving world models, this evaluation range is relatively long: DriveDreamer [53] predicts 16 future frames, GAIA-1 [20] 26 frames, Vista [13] 25 frames, and MagicDrive [12] 60 frames, whereas our setting further examines horizons up to 128 frames. Therefore, the 128-frame setting is better viewed as a stress test beyond common generation horizons; in practical horizons of around 64–80 frames (6.4–8.0s at 10Hz), LMGenDrive can still cover complete interactions such as yielding, crossing, and turning, while longer extrapolation exposes accumulated visual and control-related errors.

Table 5 shows a clear performance degradation as the prediction horizon increases, which is expected in autoregressive generation where small errors accumulate over time.

4.5 Visualization

To illustrate LMGenDrive's capabilities, Figure 4 presents qualitative videos from the CARLA simulator. The top row shows the initial multi-view observations as the conditioning inputs, while the subsequent rows visualize three future

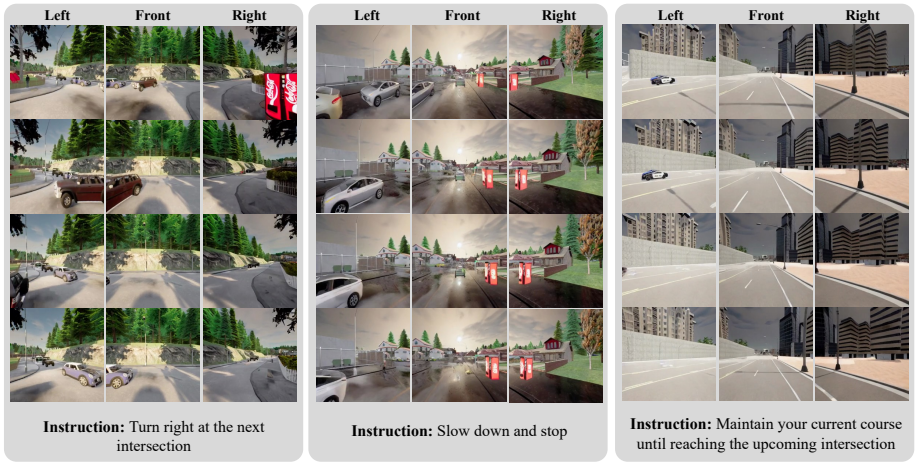

Fig. 4: Visualization of multi-view future scenes generated by LMGenDrive, showing consistent left-front-right views aligned with driving instructions.

steps generated by our multi-view world model in an autoregressive manner. At each step, the model takes the previously generated multi-view frames and predicts actions as input, to synthesize the next set of left, front, and right camera views. Each panel displays these synchronized camera views together with the corresponding driving instruction. The results show that LMGenDrive (1) preserves spatial consistency across views, (2) anticipates dynamic agents such as crossing vehicles and pedestrians, and (3) aligns future scene evolution with the given language instructions.

5 Conclusion

We introduced LMGenDrive, a unified framework that combines LLM-based multimodal reasoning with generative world models for closed-loop end-to-end driving. By jointly modeling instruction following, spatio-temporal understanding, and multi-view future video generation, the system achieves significantly improved robustness and performance on the CARLA LangAuto benchmark. Ablation studies confirm that each component—visual pretraining, action and world queries, long-horizon training, and multi-view generation—plays an essential role. Our results demonstrate the complementary strengths of integrating understanding and generation within a single architecture, enabling more anticipatory and reliable behavior than prior approaches. Looking forward, LMGenDrive provides a foundation for scaling unified reasoning–generation models to real-world settings and for exploring broader applications in embodied decision-making.

Acknowledgements

This work is supported in part by NSFC-RGC Project N_CUHK498/24, in part by Guangdong Basic and Applied Basic Research Foundation (No. 2023B1515130008, XW), in part by Shenzhen Loop Area Institute, and in part by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)'s InnoHK.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Ball, P.J., Bauer, J., Belletti, F., et al.: Genie 3: A new frontier for world models (2025)
4. CARLA Team: Carla autonomous driving leaderboard. `https://leaderboard.carla.org/` (2020), accessed: 2021-02-11
5. Chen, D., Zhou, B., Koltun, V., Krähenbühl, P.: Learning by cheating. In: Conference on Robot Learning. pp. 66–75. PMLR (2020)
6. Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., et al.: Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568 (2025)
7. Chen, L., Sinavski, O., Hünermann, J., Karnsund, A., Willmott, A.J., Birch, D., Maund, D., Shotton, J.: Driving with llms: Fusing object-level vector modality for explainable autonomous driving. arXiv preprint arXiv:2310.01957 (2023)
8. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See https://vicuna. lmsys. org (accessed 14 April 2023) **2**(3), 6 (2023)
9. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
10. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: Carla: An open urban driving simulator. In: Conference on robot learning. pp. 1–16. PMLR (2017)
11. Fu, H., Zhang, D., Zhao, Z., Cui, J., Liang, D., Zhang, C., Zhang, D., Xie, H., Wang, B., Bai, X.: Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24823–24834 (2025)
12. Gao, R., Chen, K., Xie, E., Hong, L., Li, Z., Yeung, D.Y., Xu, Q.: Magicdrive: Street view generation with diverse 3d geometry control. arXiv preprint arXiv:2310.02601 (2023)
13. Gao, S., Yang, J., Chen, L., Chitta, K., Qiu, Y., Geiger, A., Zhang, J., Li, H.: Vista: A generalizable driving world model with high fidelity and versatile controllability. arXiv preprint arXiv:2405.17398 (2024)

14. Geng, X., Gudibande, A., Liu, H., Wallace, E., Abbeel, P., Levine, S., Song, D.: Koala: A dialogue model for academic research. Blog post, April **1** (2023)
15. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
16. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
17. Ha, D., Schmidhuber, J.: World models. arXiv preprint arXiv:1803.10122 (2018)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
20. Hu, A., Russell, L., Yeo, H., Murez, Z., Fedoseev, G., Kendall, A., Shotton, J., Corrado, G.: Gaia-1: A generative world model for autonomous driving. arXiv preprint arXiv:2309.17080 (2023)
21. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., et al.: Planning-oriented autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17853–17862 (2023)
22. Jaeger, B., Chitta, K., Geiger, A.: Hidden biases of end-to-end driving models. arXiv preprint arXiv:2306.07957 (2023)
23. Ji, Y., Zhu, Z., Zhu, Z., Xiong, K., Lu, M., Li, Z., Zhou, L., Sun, H., Wang, B., Lu, T.: Cogen: 3d consistent video generation via adaptive conditioning for autonomous driving. arXiv preprint arXiv:2503.22231 (2025)
24. Jia, X., Gao, Y., Chen, L., Yan, J., Liu, P.L., Li, H.: Driveadapter: Breaking the coupling barrier of perception and planning in end-to-end autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7953–7963 (2023)
25. Jia, X., Wu, P., Chen, L., Xie, J., He, C., Yan, J., Li, H.: Think twice before driving: Towards scalable decoders for end-to-end autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21983–21994 (2023)
26. Jia, X., You, J., Zhang, Z., Yan, J.: Drivetransformer: Unified transformer for scalable end-to-end autonomous driving. arXiv preprint arXiv:2503.07656 (2025)
27. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
28. Li, Y., Fan, L., He, J., Wang, Y., Chen, Y., Zhang, Z., Tan, T.: Enhancing end-to-end autonomous driving with latent world model. arXiv preprint arXiv:2406.08481 (2024)
29. Liao, C., Liu, L., Wang, X., Luo, Z., Zhang, X., Zhao, W., Wu, J., Li, L., Tian, Z., Huang, W.: Mogao: An omni foundation model for interleaved multi-modal generation. arXiv preprint arXiv:2505.05472 (2025)
30. Liu, D., Zhang, R., Qiu, L., Huang, S., Lin, W., Zhao, S., Geng, S., Lin, Z., Jin, P., Zhang, K., et al.: Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. arXiv preprint arXiv:2402.05935 (2024)
31. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)

32. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)
33. Mao, J., Qian, Y., Ye, J., Zhao, H., Wang, Y.: Gpt-driver: Learning to drive with gpt. arXiv preprint arXiv:2310.01415 (2023)
34. Mao, J., Ye, J., Qian, Y., Pavone, M., Wang, Y.: A language agent for autonomous driving. arXiv preprint arXiv:2311.10813 (2023)
35. Qian, S., Shao, H., Zhu, Y., Li, M., Jia, J.: Blending anti-aliasing into vision transformer. Advances in Neural Information Processing Systems **34**, 5416–5429 (2021)
36. Renz, K., Chen, L., Arani, E., Sinavski, O.: Simlingo: Vision-only closed-loop autonomous driving with language-action alignment. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 11993–12003 (2025)
37. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
38. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
39. Russell, L., Hu, A., Bertoni, L., Fedoseev, G., Shotton, J., Arani, E., Corrado, G.: Gaia-2: A controllable multi-view generative world model for autonomous driving. arXiv preprint arXiv:2503.20523 (2025)
40. Sha, H., Mu, Y., Jiang, Y., Chen, L., Xu, C., Luo, P., Li, S.E., Tomizuka, M., Zhan, W., Ding, M.: Languagempc: Large language models as decision makers for autonomous driving. arXiv preprint arXiv:2310.03026 (2023)
41. Shao, H., Hu, Y., Wang, L., Song, G., Waslander, S.L., Liu, Y., Li, H.: Lmdrive: Closed-loop end-to-end driving with large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15120–15130 (2024)
42. Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. Advances in Neural Information Processing Systems **37**, 8612–8642 (2024)
43. Shao, H., Wang, L., Chen, R., Li, H., Liu, Y.: Safety-enhanced autonomous driving using interpretable sensor fusion transformer. In: Conference on Robot Learning. pp. 726–737. PMLR (2023)
44. Shao, H., Wang, L., Chen, R., Waslander, S.L., Li, H., Liu, Y.: Reasonnet: End-to-end driving with temporal and global reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13723–13733 (2023)
45. Shi, W., Han, X., Zhou, C., Liang, W., Lin, X.V., Zettlemoyer, L., Yu, L.: Lmfusion: Adapting pretrained language models for multimodal generation. arXiv preprint arXiv:2412.15188 (2024)
46. Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Beißwenger, J., Luo, P., Geiger, A., Li, H.: Drivelm: Driving with graph visual question answering. In: European Conference on Computer Vision. pp. 256–274. Springer (2024)
47. Tian, X., Gu, J., Li, B., Liu, Y., Wang, Y., Zhao, Z., Zhan, K., Jia, P., Lang, X., Zhao, H.: Drivevlm: The convergence of autonomous driving and large vision-language models. arXiv preprint arXiv:2402.12289 (2024)
48. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)

49. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
50. Wang, T., Zhang, C., Qu, X., Li, K., Liu, W., Huang, C.: Diffad: A unified diffusion modeling approach for autonomous driving. arXiv preprint arXiv:2503.12170 (2025)
51. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
52. Wang, X., Peng, P.: Prophetdwm: A driving world model for rolling out future actions and videos. arXiv preprint arXiv:2505.18650 (2025)
53. Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-drive world models for autonomous driving. In: European Conference on Computer Vision. pp. 55–72. Springer (2024)
54. Wang, Y., He, J., Fan, L., Li, H., Chen, Y., Zhang, Z.: 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 (2024)
55. Weng, X., Ivanovic, B., Wang, Y., Wang, Y., Pavone, M.: Para-drive: Parallelized architecture for real-time autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15449–15458 (2024)
56. Winter, K., Azer, M., Flohr, F.B.: Bevdriver: Leveraging bev maps in llms for robust closed-loop driving. arXiv preprint arXiv:2503.03074 (2025)
57. Xiang, J., Liu, G., Gu, Y., Gao, Q., Ning, Y., Zha, Y., Feng, Z., Tao, T., Hao, S., Shi, Y., et al.: Pandora: Towards general world model with natural language actions and video states. arXiv preprint arXiv:2406.09455 (2024)
58. Xu, Z., Zhang, Y., Xie, E., Zhao, Z., Guo, Y., Wong, K.K., Li, Z., Zhao, H.: Drivegpt4: Interpretable end-to-end autonomous driving via large language model. arXiv preprint arXiv:2310.01412 (2023)
59. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
60. Yang, J., Gao, S., Qiu, Y., Chen, L., Li, T., Dai, B., Chitta, K., Wu, P., Zeng, J., Luo, P., et al.: Generalized predictive model for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14662–14672 (2024)
61. Yang, Z., Chai, Y., Jia, X., Li, Q., Shao, Y., Zhu, X., Su, H., Yan, J.: Drivemoe: Mixture-of-experts for vision-language-action model in end-to-end autonomous driving. arXiv preprint arXiv:2505.16278 (2025)
62. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022)
63. Zhang, Z., Tang, S., Zhang, Y., Fu, T., Wang, Y., Liu, Y., Wang, D., Shao, J., Wang, L., Lu, H.: Ad-h: Autonomous driving with hierarchical agents. arXiv preprint arXiv:2406.03474 (2024)
64. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging llm-as-a-judge with mt-bench and chatbot arena (2023)
65. Zheng, Y., Liang, R., Zheng, K., Zheng, J., Mao, L., Li, J., Gu, W., Ai, R., Li, S.E., Zhan, X., et al.: Diffusion-based planning for autonomous driving with flexible guidance. arXiv preprint arXiv:2501.15564 (2025)

66. Zhou, Y., Shao, H., Wang, L., Waslander, S.L., Li, H., Liu, Y.: Smartrefine: A scenario-adaptive refinement framework for efficient motion prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15281–15290 (2024)
67. Zhou, Y., Shao, H., Wang, L., Zong, Z., Li, H., Waslander, S.L.: Drivinggen: A comprehensive benchmark for generative video world models in autonomous driving. arXiv preprint arXiv:2601.01528 (2026)
68. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
69. Zong, Z., Ma, B., Shen, D., Song, G., Shao, H., Jiang, D., Li, H., Liu, Y.: Mova: Adapting mixture of vision experts to multimodal context. Advances in Neural Information Processing Systems **37**, 103305–103333 (2024)