From Human Intention to Action Prediction: Intention-Driven End-to-End Autonomous Driving

Anonymous ACL submission

Abstract

While end-to-end autonomous driving has achieved remarkable progress in geometric control, current systems remain constrained by a command-following paradigm that relies on simple navigational instructions. Transitioning to genuinely intelligent agents requires the capability to interpret and fulfill high-level, abstract human intentions. However, this advancement is hindered by the lack of dedicated benchmarks and semantic-aware evaluation metrics. In this paper, we formally define the task of Intention-Driven End-to-End Autonomous Driving and present Intention-Drive, a comprehensive benchmark designed to bridge this gap. We construct a large-scale dataset featuring complex natural language intentions paired with high-fidelity sensor data. To overcome the limitations of conventional trajectory-based metrics, we introduce the Imagined Future Alignment (IFA), a novel evaluation protocol leveraging generative world models to assess the semantic fulfillment of human goals beyond mere geometric accuracy. Furthermore, we explore the solution space by proposing two distinct paradigms: an end-to-end vision-language planner and a hierarchical agent-based framework. The experiments reveal a critical dichotomy where existing models exhibit satisfactory driving stability but struggle significantly with intention fulfillment. Notably, the proposed frameworks demonstrate superior alignment with human intentions.

1 Introduction

End-to-end autonomous driving (E2E AD) has recently gained significant traction, promising to overcome the limitations of traditional modular pipelines by learning complex driving policies directly from data (Chen et al., 2024a; Mao et al., 2023b; Shao et al., 2024a; Wu et al., 2025). This approach has led to remarkable progress in vehicle control and navigation (Hu et al., 2023; Fu et al., 2025; Ma et al., 2024b). However, as illustrated in Figure 1(a), these systems predominantly operate at a basic intelligence level. They function as mere command-followers that execute simple navigational instructions such as "turn left", "turn right", or "go straight", rather than as intelligent agents capable of interpreting human goals (Hu et al., 2023). This fundamental limitation prevents current systems from achieving genuinely intelligent autonomy, which requires a paradigm shift from merely executing steering commands to understanding and fulfilling high-level, abstract human intentions. The leap from a command-follower to an intention-fulfiller represents a critical milestone in E2E AD intelligence, yet this transition is currently hindered by a significant gap.

The distinction between command-following and intention-fulfilling becomes apparent when examining complex real-world scenarios. As shown in Figure 1(b), a human driver can readily interpret and act upon an instruction like "parked next to the building on the right side of the road". This requires not just geometric path planning but also comprehensive scene understanding, spatial reasoning, and semantic interpretation. In contrast, as presented in Figure 1(c), conventional E2E AD systems operate on a fundamentally different architectural principle, processing simple steering commands and scene perception to generate trajectories without the comprehension of the driver's underlying goal (Hwang et al., 2024; Pan et al., 2024; Liu et al., 2025). The desired paradigm, as shown in Figure 1(d), would function as an intention-fulfiller, directly taking high-level human intention as input and requiring a comprehensive understanding that fuses intention recognition with scene perception to generate appropriate driving actions.

Recent research attempts to incorporate Large Language Models (LLMs) into autonomous driving frameworks, leveraging their semantic understanding capabilities (Han et al., 2025; Sima et al., 2024; Yuan et al., 2024; Yang et al., 2023, 2025b,a).

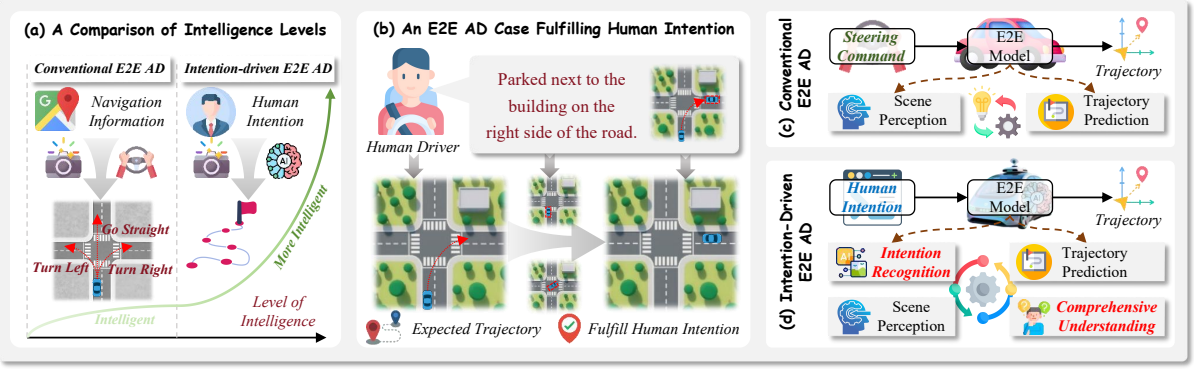


Figure 1: **Conceptual illustration of the shift from conventional to intention-driven end-to-end autonomous driving.** (a) A comparison of intelligence levels, highlighting the leap from executing low-level navigational commands to understanding high-level, abstract human intentions. (b) An example scenario where the autonomous vehicle must interpret and fulfill the complex intention, which requires reasoning beyond simple steering command. (c) The architecture of a conventional end-to-end autonomous driving system, which acts as a command-follower by translating low-level steering commands and scene perception into a trajectory. (d) Our proposed paradigm for an intention-driven system, which functions as an intention-fulfiller.

However, a critical issue has emerged: when these models undergo task-specific training using fixed, limited sets of query templates rather than diverse natural instructions, their ability to interpret human intentions degrades (Li et al., 2025a; Fu et al., 2025; Zeng et al., 2025). Although built upon LLMs that possess strong language understanding capabilities, these models become overly specialized for trajectory prediction at the expense of semantic fidelity, ultimately failing to accurately interpret complex human instructions like those shown in Figure 1(b). This phenomenon represents a significant obstacle in the development of intention-aware AD systems.

The field is further hampered by the absence of standardized evaluation frameworks that can properly measure a system's ability to understand and fulfill human intentions. Current evaluation metrics (Xu et al., 2024; Hu et al., 2023; Dauner et al., 2024) focus primarily on the geometric accuracy of trajectory prediction, which fails to capture whether the vehicle has actually satisfied the human's underlying goal. Without appropriate benchmarks to quantify this crucial capability, progress toward intention-driven autonomous driving remains unmeasurable and therefore difficult to achieve. This critical gap has persisted despite the increasing sophistication of E2E AD systems.

To address these fundamental challenges, we formulate the task of intention-driven end-to-end autonomous driving. We further introduce Intention-Drive, the first comprehensive benchmark specifically designed to evaluate an AD system's ability to translate high-level human intentions into safe and precise driving actions. Using this benchmark, we develop and evaluate two distinct methodological frameworks to tackle the intention-driven task. Through extensive experiments, we analyze the capability of current state-of-the-art models, revealing a critical dichotomy between basic driving competency and intention understanding.

The key contributions of this work are fourfold:

- We formally define *Intention-Driven End-to-End Autonomous Driving*, a new problem setting that requires models to ground high-level, abstract human intentions into driving trajectories.
- We introduce *Intention-Drive*, the first comprehensive benchmark for this task, featuring a large-scale dataset with natural language intentions and a novel evaluation protocol based on *Imagined Future Alignment (IFA)*, which assesses semantic goal fulfillment beyond geometric accuracy.
- We present two baseline frameworks: an end-to-end vision-language planner that directly maps inputs to trajectories, and a hierarchical agent-based framework that decomposes intention understanding and vehicle action.
- Our experiments uncover a fundamental disconnect between geometric driving competence and semantic intention in existing models, while our approach achieves substantially improved alignment with human intentions.

2 Related Work

End-to-end (E2E) autonomous driving replaces the traditional modular pipeline with a single, jointly optimized model that maps sensor inputs directly to driving commands or trajectories (Chen et al., 2024a; Zhang et al., 2025c). Early approaches established strong baselines by unifying perception, prediction, and planning (Hu et al., 2023; Jiang et al., 2023), while subsequent works incorporated

uncertainty modeling (Chen et al., 2024b) and generative paradigms such as diffusion models (Zheng et al., 2024; Liao et al., 2025). Despite encouraging open-loop results, E2E systems often struggle in interactive closed-loop settings, exhibiting brittleness and suboptimal behaviors (Jia et al., 2024; Fu et al., 2025). To address these limitations, recent studies increasingly explore Vision-Language-Action (VLA) models to enhance robustness and high-level decision-making (Li et al., 2025a; Zhou et al., 2025a; Chen et al., 2025). VLA models integrate multi-modal perception, language understanding, and action generation, enabling agents to interpret high-level instructions and execute grounded behaviors (Ma et al., 2024a; Driess et al., 2023). In autonomous driving, prior work emphasizes human-like cognitive structures and explicit reasoning for improved interpretability and control (Wang et al., 2025; Lu et al., 2025), including Chain-of-Thought reasoning (Yuan et al., 2025; Li et al., 2025a) and policy optimization via RL or DPO (Jiang et al., 2025; Fang et al., 2025). Recent advances further incorporate active perception to reduce uncertainty during decision-making (Zheng et al., 2025). A full version is provided in Appendix A.

3 Intention-Drive Benchmark

3.1 Task Definition

We formally define the task of intention-driven end-to-end autonomous driving. The primary objective is to generate a safe, feasible, and contextually appropriate driving trajectory that semantically satisfies a human intention. This task requires an agent to move beyond simple navigational command following and instead perform complex reasoning that grounds linguistic concepts.

Formally, let the state of the environment at timestep t be represented by a set of sensor observations O_t, consisting solely of multi-view camera images $\{I_t^1, \ldots, I_t^N\}$. Given a history of observations over a time horizon H, denoted as $\mathcal{O}_{t-H:t}$, and a high-level human intention articulated as a natural language instruction I_{lang}, the goal is to learn a policy π that predicts a future driving trajectory $T_{t+1:t+K}$. The trajectory consists of K waypoints in the bird's-eye-view (BEV) coordinate frame, $T_{t+1:t+K} = \{p_{t+1}, \ldots, p_{t+K}\}$, where each waypoint is defined as $p_k = (x_k, y_k)$. The policy is parameterized by θ and can be expressed as:

$$T_{t+1:t+K} = \pi(\mathcal{O}_{t-H:t}, I_{\text{lang}}|\theta). \quad (1)$$

Specifically, the input is defined as a tuple $(\mathcal{O}_{t-H:t}, I_{\text{lang}})$, where $\mathcal{O}_{t-H:t}$ represents a sequence of visual data capturing the history of the scene, and I_{lang} denotes a free-form natural language string describing human intention. Based on these inputs, the model predicts the output future trajectory $T_{t+1:t+K}$ for the ego-vehicle, which is formalized as a sequence of K 2D waypoints.

This task presents several fundamental challenges that distinguish it from conventional end-to-end driving. ❶ It demands robust semantic grounding, requiring the model to associate abstract linguistic concepts with their corresponding physical entities and spatial relationships in the 3D world. ❷ It necessitates sophisticated compositional reasoning to deconstruct complex instructions into a sequence of executable driving maneuvers. ❸ Evaluating success cannot rely solely on geometric metrics to a ground-truth trajectory, as multiple distinct paths could validly fulfill the same intention.

3.2 Data Construction

Existing datasets (Dauner et al., 2024; Jia et al., 2024; Caesar et al., 2019) primarily focus on trajectory prediction from navigational commands and sensor data, lacking the rich and abstract language that characterizes human intentions. To bridge this gap, we introduce Intention-Drive, a benchmark specifically designed for intention-driven end-to-end autonomous driving.

The creation of the Intention-Drive dataset follows a multi-stage pipeline designed to generate realistic and diverse intention-action pairs. Our process leverages the OpenScene dataset (Contributors, 2023) and employs advanced Large Language Models (LLMs) for language annotation. The pipeline consists of the following key steps:

Foundational Scenario Curation. The construction of our benchmark commences with the rigorous curation of driving scenarios sourced from the large-scale OpenScene dataset (Contributors, 2023; Sima et al., 2023). From this corpus, we strategically select and filter a subset of scenarios, comprising 18,765 scenes for training and 1,959 for evaluation. For each scenario, we extract a temporal sequence of sensor observations paired with the ego-vehicle's future trajectory.

Hierarchical Dataset Construction. We devised a hierarchical annotation strategy powered by an advanced Large Vision-Language Model (LVLM). Specifically, we employ GPT-5.2 as the core annotation engine, leveraging its state-of-the-art ca-

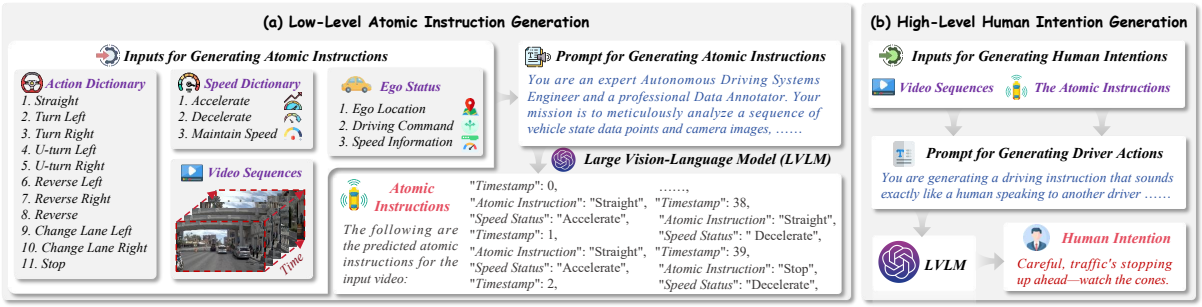


Figure 2: **The pipeline of hierarchical dataset construction.** We utilize Large Vision-Language Models (LVLMs) to annotate scenarios across three abstraction levels. (a) Low-Level Atomic Instruction Generation infers fine-grained motion primitives from vehicle states and video sequences. (b) High-Level Human Intention Generation produces natural, context-aware human intentions by reasoning over the visual scene and vehicle dynamics.

pabilities in multimodal reasoning and semantic alignment to parse complex driving scenarios. This process synthesizes two distinct levels of linguistic instructions for each driving scenario, as illustrated in Figure 2. The detailed pipeline for obtaining these instructions is described as follows:

- *Low-Level Atomic Instructions*: We first define the most granular, standardized driving primitives representing the fundamental actions of the ego vehicle. For each timestamp, a VLM is employed to infer the primary maneuver from a predefined vocabulary of atomic actions, such as `Straight`, `Turn Left`, `Change Lane Right`, and `Stop`. Concurrently, a speed status is assigned, reflecting the vehicle's instantaneous speed profile. This initial layer ensures a precise and unambiguous representation of the vehicle's movement at the most basic level, forming the foundational understanding for subsequent high-level generations.
- *High-Level Human Intentions*: Next, leveraging the comprehensive understanding derived from low-level atomic instructions, alongside the sensor inputs, these instructions capture natural language utterances from a passenger's perspective. These represent abstract goals or observations that require a higher degree of contextual reasoning and semantic interpretation. This high level focuses on mirroring the intuitive way humans communicate driving intentions, often blending direct commands with observations or warnings.

Data Verification and Refinement. To ensure annotation quality, we implement a rigorous verification pipeline where the VLM evaluates the logical coherence and semantic alignment between instruction levels. Any identified discrepancies trigger a human-in-the-loop refinement process for manual correction to guarantee the reliability.

3.3 Evaluation Protocol

To rigorously assess autonomous agents, we propose a comprehensive evaluation framework that decouples basic driving competency from high-level intention fulfillment. This protocol integrates a standardized metric for driving quality (following (Hu et al., 2023; Caesar et al., 2019)) with a novel generative approach for semantic verification.

3.3.1 Geometric Metrics

Before an agent can fulfill complex intentions, it must ensure basic driving safety and geometric stability. We employ two standard metrics to evaluate this fundamental capability:

Average Displacement Error (ADE): This metric measures the geometric fidelity of the predicted trajectory. It is calculated as the average L_2 distance between the predicted waypoints and the ground truth trajectory over the time horizon. While a low ADE indicates stable motion generation akin to the human demonstration, it is insufficient alone for intention evaluation, as multiple distinct paths may validly fulfill the same abstract intention.

Collision Rate (CR): This serves as a hard safety constraint. It is defined as the percentage of test scenarios where the ego-vehicle collides with any obstacles or road boundaries. A collision represents a critical failure of the driving system.

3.3.2 Imagined Future Alignment

While the geometric metrics ensure kinematic safety, it remains blind to the semantic nuances of human intention. To bridge this gap, we introduce Imagined Future Alignment (IFA), a novel evaluation paradigm that leverages a generative world model to explicitly reason about the future consequences of planned actions.

Conditional Future Hallucination. Instead of evaluating abstract trajectory coordinates, we map

the agent's plan back into the visual domain. We employ a pre-trained generative world model (Gao et al., 2024), denoted as $\mathcal{W}$, which serves as a neural simulator. Given the current visual observation sequence $\mathcal{O}_t$ and the agent's predicted trajectory $\mathcal{T}_{\text{pred}}$, the world model hallucinates a photo-realistic future video clip $\hat{\mathcal{V}}_{\text{future}}$:

$$\hat{\mathcal{V}}_{\text{future}} = \mathcal{W}(\mathcal{O}_t, \mathcal{T}_{\text{pred}}). \tag{2}$$

This process effectively translates the numerical driving plan into a perceptible visual narrative, enabling a more grounded semantic evaluation.

Dual-Aspect Semantic Reasoning. We utilize a VLM as a *Semantic Judge* to verify the alignment between the hallucinated future $\hat{\mathcal{V}}_{\text{future}}$ and the high-level human intention $\mathcal{I}_{\text{lang}}$. To capture both the execution quality and the final outcome, we decompose the evaluation into two distinct scores:

Process Fidelity Score ($\mathcal{S}_{\text{proc}} \in [0, 1]$): This continuous metric measures the semantic consistency of the driving behavior throughout the entire video duration. It reflects how well the agent's driving style, speed, and interaction patterns match the descriptive aspects of the intention.

Goal Completion Score ($\mathcal{S}_{\text{goal}} \in \{0, 1\}$): This binary metric assesses the success of the final state. The Judge examines the end of the hallucinated sequence to determine if the core objective (e.g., reaching a specific destination, parking in the correct slot) has been definitively achieved.

IFA Calculation. The final IFA score for a given scenario is computed as the product of the goal completion and the process fidelity. This design ensures that a scenario is only considered valid if the goal is achieved, while the score magnitude rewards precise adherence to behavioral instructions:

$$\text{IFA} = \mathcal{S}_{\text{goal}} \times \mathcal{S}_{\text{proc}} \tag{3}$$

By integrating the world model, IFA transcends traditional geometric metrics, providing a verifiable and interpretable measure of how well an autonomous agent understands and reasons about the future in the context of human commands.

4 Method

To systematically address the challenges of interpreting and executing high-level human intentions, we explore the solution space through two distinct methodological paradigms. We first introduce a unified end-to-end framework that leverages a Large Vision-Language Model to directly regress driving trajectories from multi-modal visual-linguistic inputs. Subsequently, we present a hierarchical agent-based framework that decouples the task into semantic command generation and kinematic execution, explicitly bridging the gap between abstract intention and control.

4.1 End-to-End Framework

To systematically evaluate the proposed IntentionDrive benchmark, we propose a strong end-to-end baseline. Unlike traditional E2E models that rely on simple steering commands, the intention-fulfilling task demands a model capable of deep semantic grounding, mapping abstract linguistic concepts to specific visual features and geometric actions. To this end, we adopt InternVL3.0-2B (Zhu et al., 2025) as our foundational architecture. The overall pipeline is illustrated in Figure 3(a)-(b).

Visual-Linguistic Encoding. Perceiving fine-grained scene details is a prerequisite for understanding complex human intentions. The proposed E2E framework utilizes a Vision Transformer (ViT) as the visual encoder Φ_{vis} to process the sequence of visual observations $\mathcal{O}_{1:t}$. To preserve spatial fidelity while adapting to the LLM's input space, the extracted features undergo a Pixel Unshuffle operation followed by an MLP projector. Formally, the visual tokens $\mathbf{Z}_{\text{vis}}$ are obtained as:

$$\mathbf{Z}_{\text{vis}} = \text{MLP}(\text{PixelUnshuffle}(\Phi_{\text{vis}}(\mathcal{O}_{t-H:t}))). \tag{4}$$

Simultaneously, the high-level human intention I_{lang} is processed by a text tokenizer to yield the linguistic tokens:

$$\mathbf{Z}_{\text{lang}} = \text{Tokenizer}(I_{\text{lang}}). \tag{5}$$

These visual and textual tokens are concatenated to form a unified multimodal input sequence $\mathbf{Z}_{\text{in}} = [\mathbf{Z}_{\text{vis}}, \mathbf{Z}_{\text{lang}}]$, projecting the driving scene and user intention into a shared embedding space.

LLM-Based Reasoning. The core decision-making module is powered by Qwen2.5 (Bai et al., 2025), a SoTA LLM parameterized by Θ_{E2E}. By initializing with weights from ReCogDrive (Li et al., 2025a), the model inherits a preliminary understanding of driving physics.

Trajectory Generation. We formulate the planning problem as an auto-regressive token generation task. Instead of regressing numerical values directly, the model predicts the future trajectory $T_{t+1:t+K}$ in a textual format. The trajectory

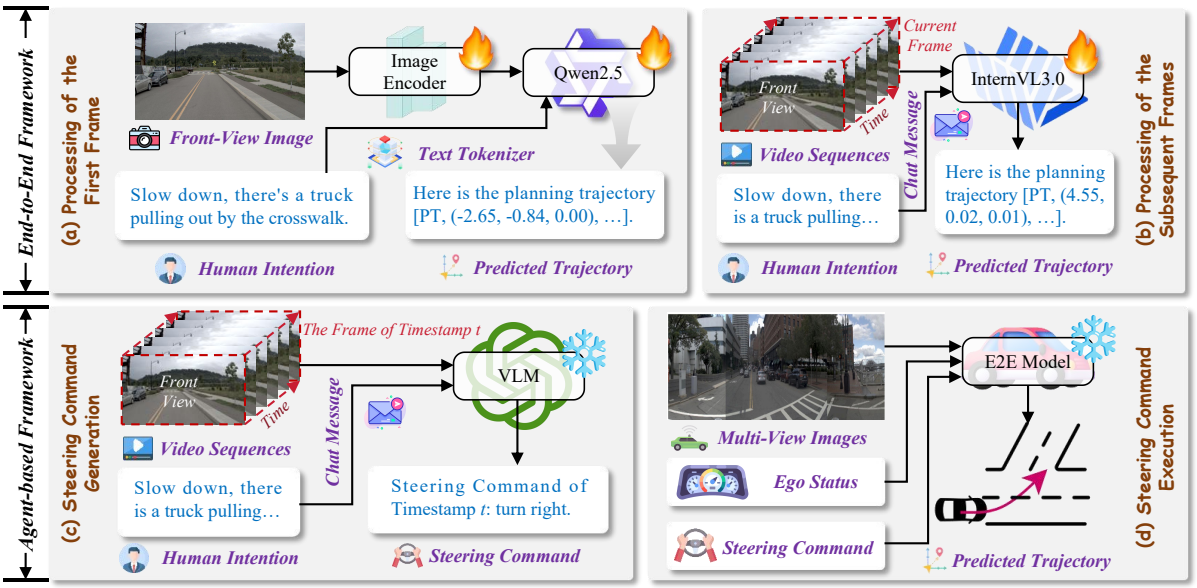


Figure 3: **Overview of the proposed methodological frameworks.** Top: The End-to-End Framework ((a)-(b)) employs InternVL3.0 to fuse visual encodings with high-level natural language intentions, directly regressing driving trajectories via a unified LLM backbone. Bottom: The Agent-Based Framework ((c)-(d)) adopts a hierarchical structure to bridge the semantic gap. A VLM first acts as a reasoning agent to translate abstract intentions into discrete steering commands, which subsequently condition a conventional E2E AD model for execution.

is serialized into a sequence of coordinate tokens $S = (s_1, s_2, \ldots, s_M)$, structured as "Here is the planning trajectory [PT, $(x_1, y_1), \ldots, (x_K, y_K)$]".

Mathematically, the model maximizes the likelihood of the target token sequence S given the multimodal inputs. The joint probability is factorized into a product of conditional probabilities:

$$P(S|\mathcal{O}_{t-H:t}, I_{\text{lang}}; \Theta_{\text{E2E}}) = \prod_{i=1}^{M} P(s_i|s_{<i}, \mathbf{Z}_{\text{vis}}, \mathbf{Z}_{\text{lang}}; \Theta_{\text{E2E}}), \quad (6)$$

where $s_{<i}$ denotes the history of generated tokens. During inference, the generated text sequence S is parsed to extract the waypoints $\{p_{t+1}, \ldots, p_{t+K}\}$ in the BEV coordinate system, constituting the final executable trajectory $T_{t+1:t+K}$.

4.2 Agent-Based Framework

Distinct from the architecture described above, the agent-based framework seeks to repurpose established conventional E2E models for the intention-driven paradigm. Since these models act primarily as command-followers conditioned on simple navigation signals, we introduce a hierarchical decomposition to bridge the semantic gap between abstract language and vehicle control. As illustrated in Figure 3(c)-(d), this framework factorizes the complex reasoning process into two sequential phases: Command Generation via a VLM agent, and Trajectory Prediction via a standard planner.

Command Generation. The objective of this phase is to translate the abstract human intention I_{lang} into a standardized steering command that acts as a deterministic interface for the downstream planner. We define a discrete action space $\mathcal{C}$, including "turn left", "turn right" and "forwrad".

We employ a Vision-Language Model (VLM) as the reasoning agent to bridge the gap. The VLM processes the sequence of visual observations $\mathcal{O}_{t-H:t}$ and the natural language intention I_{lang} to infer the optimal command c_t:

$$c_t = \arg\max_{c \in \mathcal{C}} P(c|\mathcal{O}_{t-H:t}, I_{\text{lang}}; \theta_{\text{agent}}), \quad (7)$$

where θ_{agent} represents the parameters of the reasoning agent. By explicitly predicting c_t, the system resolves the semantic ambiguity of the high-level intention before kinematic planning begins.

Trajectory Prediction. In the second phase, we leverage an off-the-shelf conventional E2E driving model, denoted as π_{E2E}. Unlike the end-to-end framework that directly regresses coordinates from language embeddings, this module operates in the traditional command-conditional mode.

Receiving the command c_t generated by the agent, along with the sensory context $\mathcal{O}_{t-H:t}$, the model generates the future trajectory $T_{t+1:t+K}$:

$$T_{t+1:t+K} = \pi_{\text{E2E}}(\mathcal{O}_{t-H:t}, c_t|\phi_{\text{drive}}), \quad (8)$$

where ϕ_{drive} represents the pre-trained weights of the driving backbone.

Method	Geometric Metrics		Intention
	ADE (m)	Coll. (%)	IFA (%)
General LLM			
GPT 5.2	1.78	0.53	23.3
Gemini 3 Flash	2.14	0.70	14.2
Gemini 3 Pro	1.58	0.42	26.8
Agent-based framework			
GPT5+RD	1.46	0.41	23.3
Gemini 3 Pro+RD	1.46	0.34	24.5
End-to-end framework			
InternVL3.0-2B	**0.70**	0.26	33.7
InternVL3.0-2B*	0.72	**0.25**	**35.6**

Table 1: **Main results on Intention-Drive.** * denotes fine-tuning initialized with weights of ReCogDrive.

Method	ADE (m)			Coll. (%)		
	1s	2s	3s	1s	2s	3s
General LLM						
GPT 5.2	0.97	1.75	2.61	0.19	0.43	0.99
Gemini 3 Flash	1.10	2.08	3.24	0.23	0.66	1.22
Gemini 3 Pro	0.51	1.41	2.83	0.19	0.35	0.71
Agent-based framework						
GPT5+RD	**0.24**	1.22	2.92	0.13	0.32	0.77
Gemini 3 Pro+RD	**0.24**	1.22	2.92	**0.09**	0.29	0.65
End-to-end framework						
InternVL3.0-2B	0.29	**0.65**	**1.16**	0.15	**0.17**	0.46
InternVL3.0-2B*	0.26	0.70	1.17	0.14	**0.17**	**0.45**

Table 2: **Performance analysis across diverse time ranges.** * denotes fine-tuning initialized with weights of ReCogDrive.

Mathematically, this agent-based approach factorizes the complex distribution of intention-driven trajectories into a chain of reasoning and execution:

$$P(T|\mathcal{O}, I_{\text{lang}}) \approx \sum_{c \in \mathcal{C}} \underbrace{P(T|\mathcal{O}, c)}_{\text{Execution}} \cdot \underbrace{P(c|\mathcal{O}, I_{\text{lang}})}_{\text{Reasoning}}. \tag{9}$$

This factorization allows us to combine the advanced semantic reasoning capabilities of VLMs with the trajectory generation of established E2E models, effectively transforming them from command-followers to intention-fulfillers.

5 Experiments

5.1 Implementation Details

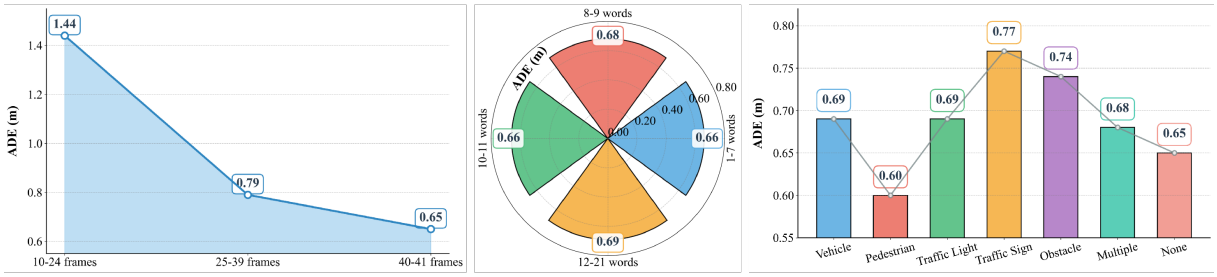
We utilize InternVL3.0-2B (Zhu et al., 2025), initializing it with the pre-trained weights from the first stage of ReCogDrive (Li et al., 2025a). The model is fine-tuned using the AdamW optimizer with a learning rate of 2×10^{-5} and a weight decay of 0.05. We employ a cosine learning rate scheduler with a warmup ratio of 0.03. The training is conducted for 1 epoch with a total batch size of 4 distributed across 4 GPUs. To process high-resolution driving scenes, we utilize a dynamic image size strategy with a base resolution of 448×448 and a maximum of 16 dynamic patches. For parameter efficiency, we apply Low-Rank Adaptation (LoRA) with a rank of $r = 32$ to both the vision backbone and the LLM. The training process is optimized using DeepSpeed with BF16 precision and gradient checkpointing.

5.2 Main Results

We evaluate all models on the Intention-Drive benchmark. Table 1 presents the quantitative results, revealing a critical dichotomy between driving competency and intention understanding.

Safety and Geometric Stability. As evidenced by the geometric metrics in Table 1, General LLMs struggle to generate kinematically feasible trajectories solely from visual-linguistic inputs. Models such as Gemini 3 Flash and GPT 5.2 exhibit high ADE of 2.14m and 1.78m, respectively, along with elevated collision rates. This indicates that while LLMs possess strong semantic reasoning, they lack the spatial grounding required to adhere to the physical constraints of driving scenes, often hallucinating unsafe paths. The agent-based frameworks improve stability slightly by decoupling reasoning from execution, yet they still lag behind the integrated approach. In contrast, the end-to-end framework demonstrates superior driving stability. The InternVL3.0-2B model achieves a remarkable ADE of 0.70m and a collision rate of 0.26%, reducing the trajectory error by approximately 55% compared to the best-performing General LLM. This confirms that end-to-end training effectively aligns high-level linguistic concepts with precise, physically grounded control policies.

Intention Fulfillment Gap. Crucially, geometric precision does not guarantee semantic alignment with human goals. The proposed IFA metric highlights this disparity. Although agent-based frameworks leverage the reasoning power of LLMs, their performance plateaus at an IFA of roughly 24%. This limitation stems from the information bottleneck inherent in the hierarchical design: compressing complex, abstract intentions into discrete steering commands results in a significant loss of semantic nuance. Conversely, the end-to-end framework establishes a direct mapping from inputs to driving trajectories, preserving the rich semantic information embedded in the human intentions. Consequently, the InternVL3.0-2B model achieves an IFA of 33.7%, which further improves to 35.6% when initialized with weights from ReCogDrive.



(a) Video sequence length (b) Intention length (c) Traffic participant

Figure 4: **Performance analysis for InternVL3.0-2B*.** (a) Analysis of planning performance across varying video sequence lengths. (b) Performance analysis of Intention lengths. (c) Performance analysis across different traffic participants.

This result empirically validates our hypothesis that a unified, intention-driven paradigm is essential for bridging the gap between understanding a command and accurately fulfilling it in complex real-world scenarios.

5.3 Analysis of Time Range

We investigate the temporal stability of the generated trajectories by decomposing the planning performance across prediction horizons of 1, 2, and 3 seconds, as shown in Table 2. While error accumulation is inevitable over longer durations due to increasing environmental uncertainty, the fine-tuned InternVL3.0-2B exhibits significantly superior robustness, maintaining a low ADE of 1.16m at the 3-second horizon. In contrast, other frameworks suffer from severe spatial drift, with greater errors. This demonstrates that our end-to-end fine-tuning strategy effectively grounds abstract intentions into precise, temporally consistent trajectory.

5.4 Analysis of Video Length

We study the impact on planning performance by categorizing test scenarios based on video sequence length, as illustrated in Figure 4(a). The results indicate that the model achieves superior performance on longer video sequences, as evidenced by the decreasing ADE. We attribute this improvement to the stabilization of planning over time: while the initial phase suffers from kinematic uncertainty, extended sequences provide richer temporal context, allowing the model to rectify early deviations and generate more consistent trajectories.

5.5 Analysis of Intention Length

To evaluate the model's robustness to linguistic complexity, we analyze the planning performance across different intention lengths, as illustrated in Figure 4(b). The results demonstrate that the model maintains a stable performance with the ADE fluctuating narrowly between 0.66m and 0.69m, regardless of the token count. This indicates that our framework effectively captures the semantic essence of high-level intentions, handling both concise commands and elaborate descriptions without significant performance degradation.

5.6 Analysis of Traffic Participants

To evaluate the model's sensitivity to specific semantic concepts, we categorize the validation scenarios based on the traffic participants explicitly referenced within the provided human intention. As illustrated in Figure 4(c), the model exhibits the lowest ADE when the intention involves dynamic agents such as "Vehicle" and "Pedestrian", benefiting from their visual saliency and predictable kinematics. In contrast, a performance drop is observed for intentions citing static regulatory elements like "Traffic Sign" and "Traffic Light", suggesting that grounding abstract linguistic instructions to small-scale, stationary visual cues requires more fine-grained reasoning than large objects.

6 Conclusion

This paper formalizes the transition in end-to-end autonomous driving from a command-following paradigm to an intention-driven intelligence level capable of fulfilling abstract human goals. To support this new task, we introduce the Intention-Drive benchmark, comprising a large-scale dataset and a generative evaluation protocol termed imagined future alignment to assess semantic goal fulfillment. We explore the solution space by proposing two distinct methodological paradigms: an end-to-end vision-language planner for direct trajectory prediction and a hierarchical agent-based framework that decouples intention reasoning from command execution. The experiments reveal a significant performance gap in existing models between driving stability and intention fulfillment, while demonstrating that our proposed frameworks achieve superior alignment with complex human instructions.

Limitations

Despite the comprehensiveness of the IntentionDrive benchmark, the current dataset is primarily focused on typical urban driving environments, which may not fully capture the complex dynamics of extreme weather conditions or highly irregular rural road layouts. Future research will focus on expanding the diversity of the benchmark to include more varied geographical and environmental contexts, while exploring the integration of multi-agent social reasoning to further enhance the alignment between human intentions and driving actions.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2019. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*.

Wei Cao, Marcel Hallgarten, Tianyu Li, Daniel Dauner, Xunjiang Gu, Caojun Wang, Yakov Miron, Marco Aiello, Hongyang Li, Igor Gilitschenski, and 1 others. 2025. Pseudo-simulation for autonomous driving. *arXiv preprint arXiv:2506.04218*.

Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. 2024a. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Shaoyu Chen, Bo Jiang, Hao Gao, Bencheng Liao, Qing Xu, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. 2024b. Vadv2: End-to-end vectorized autonomous driving via probabilistic planning. *arXiv preprint arXiv:2402.13243*.

Yuntao Chen, Yuqi Wang, and Zhaoxiang Zhang. 2025. Drivinggpt: Unifying driving world modeling and planning with multi-modal autoregressive transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

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

Yaodong Cui, Shucheng Huang, Jiaming Zhong, Zhenan Liu, Yutong Wang, Chen Sun, Bai Li, Xiao Wang, and Amir Khajepour. 2023. Drivellm: Charting the path toward full autonomous driving with large language models. *IEEE Transactions on Intelligent Vehicles*.

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

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. Carla: An open urban driving simulator. In *Conference on robot learning*.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, and 1 others. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Shiyu Fang, Yiming Cui, Haoyang Liang, Chen Lv, Peng Hang, and Jian Sun. 2025. Corevla: A dual-stage end-to-end autonomous driving framework for long-tail scenarios via collect-and-refine. *arXiv preprint arXiv:2509.15968*.

Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Dingkang Liang, Chong Zhang, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. 2025. Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. *arXiv preprint arXiv:2503.19755*.

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

Wencheng Han, Dongqian Guo, Cheng-Zhong Xu, and Jianbing Shen. 2025. Dme-driver: Integrating human decision logic and 3d scene perception in autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, and 1 others. 2023. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, and 1 others. 2024. Emma: End-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2410.23262*.

Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. 2024. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. *Advances in Neural Information Processing Systems*.

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

Bo Jiang, Shaoyu Chen, Qian Zhang, Wenyu Liu, and Xinggang Wang. 2025. Alphadrive: Unleashing the power of vlms in autonomous driving via reinforcement learning and reasoning. *arXiv preprint arXiv:2503.07608*.

Yongkang Li, Kaixin Xiong, Xiangyu Guo, Fang Li, Sixu Yan, Gangwei Xu, Lijun Zhou, Long Chen, Haiyang Sun, Bing Wang, and 1 others. 2025a. Recogdrive: A reinforced cognitive framework for end-to-end autonomous driving. *arXiv preprint arXiv:2506.08052*.

Yue Li, Meng Tian, Dechang Zhu, Jiangtong Zhu, Zhenyu Lin, Zhiwei Xiong, and Xinhai Zhao. 2025b. Drive-r1: Bridging reasoning and planning in vlms for autonomous driving with reinforcement learning. *arXiv preprint arXiv:2506.18234*.

Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, and 1 others. 2025. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*.

Haochen Liu, Zhiyu Huang, Wenhui Huang, Haohan Yang, Xiaoyu Mo, and Chen Lv. 2025. Hybrid-prediction integrated planning for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yuhang Lu, Jiadong Tu, Yuexin Ma, and Xinge Zhu. 2025. Real-ad: Towards human-like reasoning in end-to-end autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Yuechen Luo, Fang Li, Shaoqing Xu, Zhiyi Lai, Lei Yang, Qimao Chen, Ziang Luo, Zixun Xie, Shengyin Jiang, Jiaxin Liu, and 1 others. 2025. Adathinkdrive: Adaptive thinking via reinforcement learning for autonomous driving. *arXiv preprint arXiv:2509.13769*.

Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. 2024a. A survey on vision-language-action models for embodied ai. *arXiv preprint arXiv:2405.14093*.

Yunsheng Ma, Xu Cao, Wenqian Ye, Can Cui, Kai Mei, and Ziran Wang. 2024b. Learning autonomous driving tasks via human feedbacks with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. 2023a. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*.

Jiageng Mao, Junjie Ye, Yuxi Qian, Marco Pavone, and Yue Wang. 2023b. A language agent for autonomous driving. *arXiv preprint arXiv:2311.10813*.

Mona Mirzaie and Bodo Rosenhahn. 2025. Interpretable decision-making for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar, and Liu Ren. 2024. Vlp: Vision language planning for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*.

Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. 2024a. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024b. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. 2024. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*.

Chonghao Sima, Wenwen Tong, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, and Hongyang Li. 2023. Scene as occupancy.

Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. 2025. Sparsedrive: End-to-end autonomous driving via sparse scene representation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*.

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

Zhennan Wang, Jianing Teng, Canqun Xiang, Kangliang Chen, Xing Pan, Lu Deng, and Weihao Gu. 2025. Cogad: Cognitive-hierarchy guided end-to-end autonomous driving. *arXiv preprint arXiv:2505.21581*.

Yaozu Wu, Dongyuan Li, Yankai Chen, Renhe Jiang, Henry Peng Zou, Wei-Chieh Huang, Yangning Li, Liancheng Fang, Zhen Wang, and Philip S Yu. 2025. Multi-agent autonomous driving systems with large language models: A survey of recent advances, resources, and future directions. *Findings of the Association for Computational Linguistics: EMNLP*.

Yichen Xie, Runsheng Xu, Tong He, Jyh-Jing Hwang, Katie Luo, Jingwei Ji, Hubert Lin, Letian Chen, Yiren Lu, Zhaoqi Leng, and 1 others. 2025. S4-driver: Scalable self-supervised driving multimodal large language model with spatio-temporal visual representation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*.

Chengkai Xu, Jiaqi Liu, Shiyu Fang, Yiming Cui, Dong Chen, Peng Hang, and Jian Sun. 2025. Tell-drive: Enhancing autonomous driving with teacher llm-guided deep reinforcement learning. *arXiv preprint arXiv:2502.01387*.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. 2024. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*.

Tianyi Yan, Wencheng Han, Xia Zhou, Xueyang Zhang, Kun Zhan, Cheng-zhong Xu, and Jianbing Shen. 2025. Rlgf: Reinforcement learning with geometric feedback for autonomous driving video generation. *arXiv preprint arXiv:2509.16500*.

Kairui Yang, Zihao Guo, Gengjie Lin, Haotian Dong, Zhao Huang, Yipeng Wu, Die Zuo, Jibin Peng, Ziyuan Zhong, Xin Wang, and 1 others. 2025a. Trajectory-llm: A language-based data generator for trajectory prediction in autonomous driving. In *The Thirteenth International Conference on Learning Representations*.

Zhenjie Yang, Yilin Chai, Xiaosong Jia, Qifeng Li, Yuqian Shao, Xuekai Zhu, Haisheng Su, and Junchi Yan. 2025b. Drivemoe: Mixture-of-experts for vision-language-action model in end-to-end autonomous driving. *arXiv preprint arXiv:2505.16278*.

Zhenjie Yang, Xiaosong Jia, Hongyang Li, and Junchi Yan. 2023. Llm4drive: A survey of large language models for autonomous driving. *arXiv preprint arXiv:2311.01043*.

Jianhao Yuan, Shuyang Sun, Daniel Omeiza, Bo Zhao, Paul Newman, Lars Kunze, and Matthew Gadd. 2024. Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model. *arXiv preprint arXiv:2402.10828*.

Zhenlong Yuan, Jing Tang, Jinguo Luo, Rui Chen, Chengxuan Qian, Lei Sun, Xiangxiang Chu, Yujun Cai, Dapeng Zhang, and Shuo Li. 2025. Autodrive-r^2: Incentivizing reasoning and self-reflection capacity for vla model in autonomous driving. *arXiv preprint arXiv:2509.01944*.

Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, Xing Wei, and Ning Guo. 2025. Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. *arXiv preprint arXiv:2505.17685*.

Jiawei Zhang, Xuan Yang, Taiqi Wang, Yu Yao, Aleksandr Petiushko, and Bo Li. 2025a. Safeauto: Knowledge-enhanced safe autonomous driving with multimodal foundation models. *arXiv preprint arXiv:2503.00211*.

Ruifei Zhang, Junlin Xie, Wei Zhang, Weikai Chen, Xiao Tan, Xiang Wan, and Guanbin Li. 2025b. Adadrive: Self-adaptive slow-fast system for language-grounded autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Ruifei Zhang, Wei Zhang, Xiao Tan, Sibei Yang, Xiang Wan, Xiaonan Luo, and Guanbin Li. 2025c. Vldrive: Vision-augmented lightweight mllms for efficient language-grounded autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Songyan Zhang, Wenhui Huang, Zhan Chen, Chua Jiahao Collister, Qihang Huang, and Chen Lv. 2025d. Openread: Reinforced open-ended reasoing for end-to-end autonomous driving with llm-as-critic. *arXiv preprint arXiv:2512.01830*.

Weicheng Zheng, Xiaofei Mao, Nanfei Ye, Pengxiang Li, Kun Zhan, Xianpeng Lang, and Hang Zhao. 2025. Driveagent-r1: Advancing vlm-based autonomous driving with active perception and hybrid thinking. *arXiv preprint arXiv:2507.20879*.

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

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

Zewei Zhou, Tianhui Cai, Seth Z Zhao, Yun Zhang, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. 2025b. Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. *arXiv preprint arXiv:2506.13757*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Related Work

A.1 End-to-end Autonomous Driving

End-to-end (E2E) autonomous driving has emerged as a dominant paradigm, aiming to replace the traditional modular pipeline of perception, prediction, and planning with a single, jointly optimized neural network (Chen et al., 2024b,a; Zhang et al., 2025a; Xie et al., 2025; Zhang et al., 2025c; Mirzaie and Rosenhahn, 2025). This approach directly maps raw sensor inputs, such as images and LiDAR point clouds, to driving commands or future trajectories, thereby minimizing error accumulation across cascaded modules. Pioneering works in this domain have demonstrated the efficacy of integrating multiple intermediate tasks into a unified, planning-oriented framework (Chen et al., 2024a; Sun et al., 2025; Yan et al., 2025). For instance, UniAD (Hu et al., 2023) and VAD (Jiang et al., 2023) established strong baselines by holistically modeling the driving scene and formulating the entire system as a planning-centric problem. These methods showcase impressive performance by leveraging shared representations across perception and prediction to benefit the final planning output. However, deterministic trajectory regression fails to capture the inherent uncertainty and multi-modal nature of real-world driving scenarios. To address this, recent efforts have shifted towards modeling a distribution over possible future actions. VADv2 (Chen et al., 2024b) introduced probabilistic planning by predicting a probability distribution over a set of actions and sampling from it to control the vehicle. Concurrently, a new paradigm employing generative models has gained traction. Methods like GenAD (Zheng et al., 2024) and DiffusionDrive (Liao et al., 2025) leverage the power of diffusion models to generate diverse and plausible multi-modal trajectories, better reflecting the complex decision-making process of human drivers.

Despite promising results on open-loop benchmarks, where models predict trajectories on prerecorded logs, a significant performance gap often exists when these systems are deployed in interactive, closed-loop simulations. As highlighted by recent studies (Jia et al., 2024; Cao et al., 2025), many E2E models tend to overfit to the ego-vehicle's status in the training data, leading to suboptimal or unsafe behaviors in dynamic, reactive environments (Fu et al., 2025). While some works adopt closed-loop evaluation in simulators like CARLA (Dosovitskiy et al., 2017), their performance often reveals the brittleness of models trained primarily for open-loop metrics. This discrepancy underscores the critical need for developing E2E frameworks that are not only accurate in offline evaluation but also robust and reliable in realistic, interactive driving scenarios. To address these limitations, several recent approaches have successfully employed Vision-Language-Action (VLA) models, demonstrating strong closed-loop performance in autonomous driving systems (Fu et al., 2025; Li et al., 2025a; Zhou et al., 2025a; Chen et al., 2025).

A.2 Vision-Language-Action Model

Vision-Language-Action (VLA) models represent a paradigm shift in creating intelligent agents, aiming to bridge the gap between multi-modal perception, natural language understanding, and physical action generation in the real world (Ma et al., 2024a; Zhou et al., 2025b; Li et al., 2025b). Originating from the field of robotics, VLAs are designed to interpret high-level instructions and execute them by generating a sequence of low-level actions, making them a natural architectural choice for intention-driven autonomous driving (Driess et al., 2023; Tian et al., 2024; Mao et al., 2023a; Zhang et al., 2025b,d; Cui et al., 2023). These models move beyond passive scene description and towards active interaction, which is critical for fulfilling complex human intentions in dynamic environments like driving.

In the context of autonomous driving, a significant body of work aims to emulate human-like cognitive structures to make decision-making more robust and interpretable (Xu et al., 2025). For instance, models like CogAD (Wang et al., 2025) and ReAL-AD (Lu et al., 2025) explicitly design hierarchical frameworks inspired by cognitive psychology, breaking down the planning process from high-level intention or strategy to fine-grained trajectory execution. This structured approach is often enhanced by sophisticated reasoning mechanisms. AutoDrive-R2 (Yuan et al., 2025) incorporates a chain-of-thought process with self-reflection, while ReCogDrive (Li et al., 2025a) integrates a cognitive VLM with a diffusion planner to address the modality mismatch between discrete language commands and continuous driving actions. To move beyond the limitations of simple imitation learning from static datasets, reinforcement learning (RL) has emerged as a key technique for policy refinement. AlphaDrive (Jiang et al., 2025) leverages

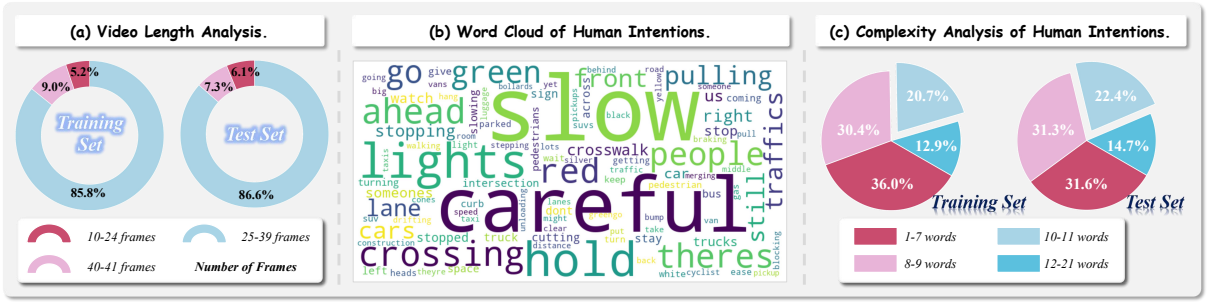


Figure 5: **Statistical analysis of the Intention-Drive benchmark.** (a) Distribution of video frame counts, ensuring coverage of diverse temporal horizons. (b) Word cloud of human intentions illustrating semantic richness. (c) Distribution of instruction lengths, highlighting the linguistic complexity with a significant portion of long, detailed intentions.

Group Relative Policy Optimization (GRPO) (Shao et al., 2024b) with planning-oriented rewards to significantly boost performance. Similarly, AdaThinkDrive (Luo et al., 2025) employs RL to train an agent that can adaptively switch between "fast" direct prediction and "slow" deliberative reasoning based on scene complexity. Other approaches focus on learning from more nuanced human feedback; CoReVLA (Fang et al., 2025), for example, uses Direct Preference Optimization (DPO) (Rafailov et al., 2023) to refine its model from sparse human takeover data in long-tail scenarios. Pushing the frontier even further, DriveAgent-R1 (Zheng et al., 2025) pioneers the concept of active perception, where the agent can proactively use vision tools to seek out additional information to resolve uncertainty, representing a paradigm shift from passive observation to active, grounded decision-making.

B Dataset Analysis

To better understand the challenges posed by Intention-Drive, we provide a comprehensive statistical analysis of the dataset characteristics, as illustrated in Figure 5. The analysis focuses on three key dimensions: temporal diversity, semantic richness, and linguistic complexity.

Temporal Diversity. Figure 5(a) presents the distribution of video sequence lengths across the training and test sets. The dataset covers a broad range of temporal horizons, reflecting the variable duration of real-world driving maneuvers. The distribution indicates that the dataset is not limited to short-term reactions but includes extended sequences that require the agent to maintain temporal consistency over longer periods.

Semantic Richness. Unlike traditional datasets restricted to steering commands, our dataset emphasizes the behavioral and interactive aspects of driving. The word cloud in Figure 5(b) highlights high-frequency keywords such as "slow", "careful", "traffic", "wait", and "pedestrian". The prominence of these terms underscores the shift from command-following to intention-fulfillment.

Linguistic Complexity. Figure 5(c) analyzes the complexity of human intentions based on instruction length. Unlike conventional datasets dominated by short phrases, our benchmark features a substantial distribution of long-form instructions. Specifically, instructions exceeding 7 words constitute the majority. This linguistic richness necessitates strong language grounding capabilities, as the E2E AD model must parse the sentence to extract the destination, behavioral modifiers, and safety constraints embedded in the driver's intention.

C Case Study

To qualitatively validate the effectiveness of our framework and the proposed Imagined Future Alignment (IFA) protocol, we present a visualization of the generative evaluation process in Figure 6. This figure demonstrates how the pre-trained generative world model "hallucinates" the future consequences of the agent's predicted trajectory, enabling a direct semantic comparison against the ground truth (Real Future) and the user's high-level instruction. We select two representative scenarios to illustrate the model's capability in handling safety-critical constraints and specific navigational goals.

In the left panel, the user issues a complex, safety-oriented command: *Wait, the light's still red and that car's cutting across — hold on.* This instruction requires the agent to prioritize immediate safety over progress and recognize the dynamic hazard. The visualized "Fake Future" confirms that the agent successfully generated a stationary trajectory. Crucially, because the trajectory was correct, the world model renders the crossing vehicle passing safely in front of the ego-vehicle, mirroring the dynamics observed in the "Real Future." The Se-

Figure 6: **Case study for InternVL3.0-2B*.** We display the historical context, the ground truth future (Real Future), and the hallucinated future (Fake Future) generated by the world model conditioned on our agent's predicted trajectory. The comparison highlights that the generated "Fake Future" successfully reflects the semantic constraints of the human instructions—such as holding for a cutting vehicle (left) or identifying a specific valet entrance (right)—thereby verifying the semantic consistency of the planned action beyond simple geometric metrics.

mantic Judge (VLM) evaluating this hallucinated clip can thus verify that the "hold on" condition was met, resulting in a high Score.

Conversely, the right panel depicts a goal-oriented scenario where the driver instructs: *Turn right into the valet and watch for pedestrians.* The challenge here lies in grounding the semantic entity "valet" to the correct spatial location rather than a generic road intersection. The "Fake Future" visualization shows the perspective shifting smoothly into the specific driveway, closely aligning with the visual cues in the ground truth video. This demonstrates that the agent's planned trajectory effectively guided the world model to render the correct destination. These visualizations highlight that our Intention-Driven framework does not merely regress coordinates but semantically fulfills the human's underlying intent, which is faithfully captured and verified by the IFA protocol.