ReasonPlan: Unified Scene Prediction and Decision Reasoning for Closed-loop Autonomous Driving

Xueyi Liu[1,2,3] Zuodong Zhong[4] Junli Wang[1,2] Yuxin Guo[1,2] Zhiguo Su[3]
Qichao Zhang[1,2]* Yun-Fu Liu[3] Yinfeng Gao[4] Yupeng Zheng[1,2] Qiao Lin[3]
Huiyong Chen[3] Dongbin Zhao[1,2]*

[1]SKL-MAIS, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]EACON, Fujian, China
[4]School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing, China

Abstract: Due to the powerful vision-language reasoning and generalization abilities, multimodal large language models (MLLMs) have garnered significant attention in the field of end-to-end (E2E) autonomous driving. However, their application to closed-loop systems remains underexplored, and current MLLM-based methods have not shown clear superiority to mainstream E2E imitation learning approaches. In this work, we propose ReasonPlan, a novel MLLM fine-tuning framework designed for closed-loop driving through holistic reasoning with a self-supervised Next Scene Prediction task and supervised Decision Chain-of-Thought process. This dual mechanism encourages the model to align visual representations with actionable driving context, while promoting interpretable and causally grounded decision making. We curate a planning-oriented decision reasoning dataset, namely PDR, comprising 210k diverse and high-quality samples. Our method outperforms the mainstream E2E imitation learning method by a large margin of 16.44% L2 and 16.1 driving score on Bench2Drive benchmark. Furthermore, ReasonPlan demonstrates strong zero-shot generalization on unseen DOS benchmark, highlighting its adaptability in handling zero-shot corner cases. Code and dataset will be found in `https://github.com/Liuxueyi/ReasonPlan`.

Keywords: Multimodal LLM, Closed-Loop Evaluation, Autonomous Driving

1 Introduction

Recently, end-to-end (E2E) autonomous driving presents a scalable, data-driven paradigm that has garnered increasing attention [1, 2, 3]. Despite its advantages in simplifying the driving pipeline, most existing E2E approaches rely on imitation learning [4, 5] and exhibit limitations in complex, closed-loop environments. Specifically, they often suffer from causal confusion during interactive cases [6] and struggle to generalize to out-of-distribution scenarios [7]. Recent progress in multimodal large language models (MLLMs) [8, 9, 10] enables vision-language reasoning [11] and zero-shot generalization [12] capabilities, offering new opportunities for E2E autonomous driving.

Recent efforts have explored dual-system frameworks [13, 14, 15], LLM distillation for enhancing E2E driving [16, 17], and direct trajectory prediction in textual form [18, 19, 20]. While promising, these approaches predominantly operate in open-loop settings or exhibit suboptimal performance in closed-loop evaluations. This limitation stems from their inability to perform context-aware reasoning and robust planning in closed-loop scenarios, where continuous adaptation to dynamic environments is essential [21]. We conclude three key challenges that limit the full exploitation of MLLMs'

*Corresponding author

9th Conference on Robot Learning (CoRL 2025), Seoul, Korea.

reasoning capabilities: (1) **Insufficient utilization of visual information.** Existing methods [22] supervise the perception and decision-making process using only text or simply use predicted images as input for action generation [23], leading to constrained scene comprehension and the accumulation of compounded errors. (2) **No explicit reasoning process.** Prior works [18, 19] utilize multi-turn QAs fine-tuning LLMs to enhance the instruction-following ability, but they fall short of engaging the models' Chain-of-Thought (CoT) reasoning capabilities. (3) **Lack of planning-oriented high-quality reasoning datasets.** Consequently, the full potential of MLLMs for E2E closed-loop planning remains underexplored, and current MLLM-based methods have yet to demonstrate clear superiority over imitation learning techniques [1, 2, 24] in closed-loop benchmarks.

To address the challenges above, we explore both model architecture and training datasets. From the model perspective, we propose a novel MLLM fine-tuning framework, namely **ReasonPlan**, which effectively enhances the holistic reasoning capabilities in complicated closed-loop scenarios. ReasonPlan comprises a **self-supervised Next Scene Prediction (NSP) task** and **supervised Decision Chain-of-Thought (DeCoT) process**. Specifically, motivated by recent advances in scene understanding and generative modeling [25, 26], we introduce a temporally NSP task that conditions on ego context to forecast future visual observations. This objective constrains image token representations in the latent space, enriching the model's contextual understanding of driving scenarios. To further leverage the reasoning and generalization capabilities under closed-loop and out-of-distribution evaluations, we incorporate explicit textual supervision over the DeCoT process.

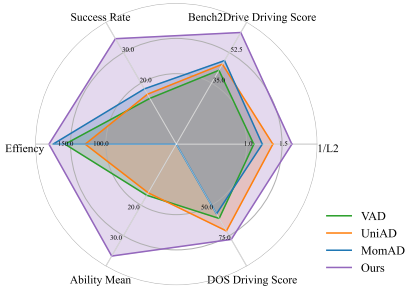


Figure 1: The proposed ReasonPlan achieves leading performance on most of metrics compared with E2E methods.

From the dataset perspective, we construct a large-scale instruction dataset tailored for closed-loop planning, called **PDR**, which contains 203,353 training samples and 11,047 testing samples. Using an automated annotation pipeline, PDR captures the complete decision reasoning process in training scenarios on the Bench2Drive [27], including the following stages: *Scene Understanding*, *Traffic Sign Recognition*, *Critical Object Identification for Risk Assessment*, and *Meta Action*. The dataset will be released publicly, serving as a foundation for learning structured and causally grounded decision reasoning.

ReasonPlan achieves a **driving score of 64.01** and reduces the **L2 error by 16.44%** on Bench2Drive [27], outperforming prior state-of-the-art E2E imitation learning models. Furthermore, it demonstrates **strong zero-shot generalization** on DOS [28], highlighting its adaptability in decision-critical corner cases.

Our key contributions are as follows. (1) We propose ReasonPlan, a novel MLLM fine-tuning framework for complex closed-loop driving scenarios. The integration of NSP and DeCoT effectively couples the visual and language modalities, enabling a comprehensive decision reasoning process. (2) We construct PDR, a large-scale decision reasoning dataset via an automated annotation pipeline tailored for closed-loop planning, containing 210k diverse and high-quality samples. (3) As shown in Fig. 1, ReasonPlan demonstrates superior performance on the Bench2Drive under both open-loop and closed-loop settings, and shows strong zero-shot generalization on DOS scenarios.

2 Related Work

End-to-End Autonomous Driving. Current mainstream E2E driving directly maps raw sensor inputs to trajectories based on imitation learning (IL). Representative works such as UniAD [1] and VAD [2] unify perception, prediction, and planning into a single framework, enabling joint optimization across the full pipeline. To address planning uncertainty, SparseDrive [5] adopts sparse representations in a multimodal planning framework, while UncAD [3] integrates uncertainty-aware online mapping. VADv2 [4] models a probabilistic distribution over actions, enhancing robustness via action sampling. However, these methods are evaluated in open-loop settings, where the models

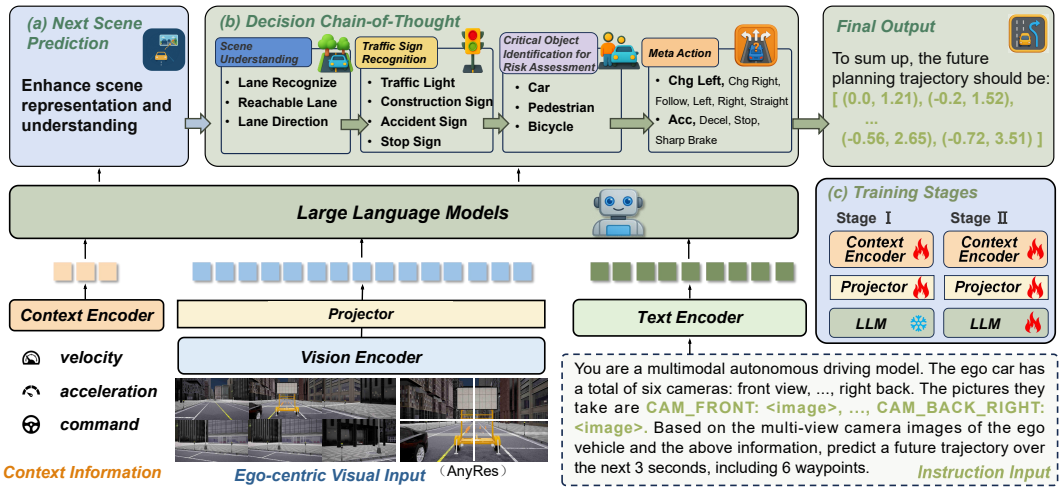


Figure 2: The pipeline of ReasonPlan, a holistic reasoning framework for closed-loop driving. It consists of two main modules: (a) the next scene prediction to enhance scene representation and understanding, which is conditioned on current context information; (b) the supervised decision CoT process to obtain the final planning trajectory. (c) the two training stages.

tend to overfit to specific ego-vehicle states [29]. DiffusionDrive [30] and GoalFlow [31] explore a novel generative paradigm, leveraging diffusion models to predict diverse future trajectories in Navsim [32]. Some other studies [33, 24] have adopted closed-loop evaluation in CARLA[34] to assess driving robustness. However, those IL-based E2E still suffer from significant causal confusion [6] and limited generalization capabilities [7]. To this end, we propose a MLLM-based E2E fine-tuning framework, aiming to harness their pretrained world knowledge and reasoning capabilities to address the challenges of closed-loop driving.

Multimodal Large Language Models for Autonomous Driving. MLLMs have exhibited impressive capabilities in scene understanding and high-level reasoning across language, vision, and robotics domains [35, 36, 37, 38, 39], motivating their integration into autonomous driving systems. Recent studies incorporate MLLMs into E2E frameworks via dual-system architectures and knowledge distillation [14, 15, 16, 17]. DriveVLM [14] and Senna [15] leverage MLLMs to produce high-level driving intentions, which are subsequently refined by low-level policy modules for final trajectory generation. VLP [16] and DiMA [17] align key components of E2E systems with MLLMs, distilling abstract reasoning capabilities into lightweight planning heads. Other methods [20, 19] adopt simple QAs fine-tuning to generate textual trajectories. For example, Reason2Drive [20] enhances scene comprehension through CoT datasets and structured tokenization, while TOKEN [19] improves long-tail planning by combining object-level perception with LLM-based reasoning. Despite promising results, these approaches are limited to open-loop evaluations. While some methods explore closed-loop settings [18, 13, 21], they often rely on simplified benchmarks such as Town05Long [40] or HighwayEnv [41]. LMDrive [18] introduces a language-based closed-loop framework for autonomous driving, lacking structured reasoning tasks. SimLingo [42], built upon CarLLaVA [43], introduces an action dreaming task to bridge language and control action spaces. To fully exploit the reasoning potential of MLLMs in complex interactive scenarios, we propose a unified framework that tightly integrates visual and textual modalities, enabling comprehensive decision reasoning and zero-shot generalization in closed-loop scenarios.

3 Method

Overview. The overall pipeline of ReasonPlan is illustrated in Fig. 2. Specifically, ReasonPlan comprises two components: (a) the self-supervised Next Scene Prediction task (Sec. 3.1), which aims to enhance scene representation and understanding; (b) the supervised Decision Chain-of-Thought process (Sec. 3.2) for reasoning and interpretable planning. Moreover, (c) the framework is optimized in two stages (Sec. 3.3).

3.1 Self-supervised Next Scene Prediction (NSP)

To enhance scene understanding and effectively align the visual feature space with the semantic space of language, we introduce a temporally self-supervised NSP task, as illustrated in Fig. 3. The input to our model consists of multi-view RGB images denoted as $\mathbf{X}_{v_t} \in \mathbb{R}^{N \times W \times H \times 3}$, where N is the number of views, and W and H are the width and height of each image. We first resize the all-view images into grids. Additionally, we adopt an AnyRes [44] partitioning strategy that divides the front view $\mathbf{X}_{f_t}$ into four spatial grids to increase the model's efficiency and ability to capture fine-grained spatial details. Each gird is processed by the vision encoder SigLIP [45], yielding a visual feature tensor $\mathbf{Z}_{v_t} \in \mathbb{R}^{L_v \times D_v}$, where L_v and D_v denote the number of visual tokens per grid and the visual embedding dimension. To align these features with the textual space, we apply a two-layer MLP projection module that maps $\mathbf{Z}_{v_t}$ to $\mathbf{H}_{v_t} \in \mathbb{R}^{L_v \times D_p}$, where D_p denotes the language embedding dimension. This process can be formally written as:

$$\mathbf{Z}_{v_t} = \texttt{SigLIP}(\texttt{AnyRes}(\mathbf{X}_{v_t})), \quad \mathbf{H}_{v_t} = \texttt{MLP}(\mathbf{Z}_{v_t}), \tag{1}$$

Considering the critical role of both vehicle dynamics and high-level driving intent, we introduce a context encoder, a two-layer MLP module, to embed the ego vehicle's current velocity $\boldsymbol{v}$, acceleration $\boldsymbol{a}$, and driving command `cmd` (e.g. turn left) into a context representation:

$$\mathbf{H}_{c_t} = \texttt{MLP}(\boldsymbol{v}, \boldsymbol{a}, \texttt{cmd}), \tag{2}$$

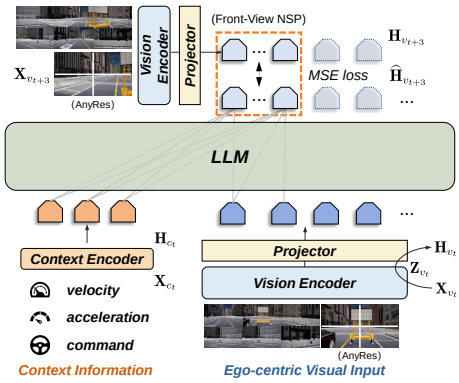


Figure 3: The process of NSP task.

This encoded context is then fused with the visual features $\mathbf{H}_{v_t}$ to guide the prediction of future scene representation. After the LLM, we can estimate the latent visual embedding $\hat{\mathbf{H}}_{v_{t+3}}$. For self-supervised labeling, we utilize the multi-view images at a 3-second time horizon $\mathbf{X}_{v_{t+3}}$ as the target future frame. These images are processed through the identical SigLIP encoder and MLP projection after AnyRes grid processing, yielding latent visual representations $\mathbf{H}_{v_{t+3}}$ as self-supervision signals. In autonomous driving, the front-view image contains the most semantically informative content, capturing critical cues for trajectory planning. Our ablation study (in Appendix D) also reveals that, for NSP task, training front-view achieves performance on par with the full-view setting. To improve training efficiency and eliminate redundant computation, we uniformly adopt the front-view NSP in subsequent experiments. While the MSE loss is computed solely on the front-view, the full image features are retained as condition, providing comprehensive contextual information for the subsequent reasoning process.

$$\hat{\mathbf{H}}_{v_{t+3}} = \texttt{LLM}(\texttt{Concat}(\mathbf{H}_{c_t}, \mathbf{H}_{v_t})), \quad \mathcal{L}_{\text{image}} = \|\hat{\mathbf{H}}_{v_{t+3}}[:\texttt{front}] - \mathbf{H}_{v_{t+3}}[:\texttt{front}]\|^2, \tag{3}$$

We employ a Mean Squared Error (MSE) loss in latent feature space to enforce consistency between predicted and future visual representations. By conditioning on both dynamic and contextual cues, our model anticipates future perceptual states aligned with intended maneuvers. Through a temporally self-supervised NSP task, it integrates ego motion and scene context to enhance the LLM's spatial awareness, enabling comprehensive scene understanding and context-aware forecasting.

3.2 Decision Chain-of-Thought (DeCoT)

Following the NSP task, the DeCoT process in the textual space is pivotal in our framework. Reason-Plan leverages the pretrained knowledge of LLM, integrating visual tokens and textual instructions to generate interpretable decision-making processes and produce executable trajectories in text form.

As shown in Fig. 2, the system prompt and navigation instructions $\mathbf{X}_p$ are first tokenized into token IDs and then encoded into textual embeddings $\mathbf{H}_p^{L_p \times D_p}$ using a text encoder, where L_p denotes the number of text tokens. To enable effective multimodal alignment, we introduce a special `<image>` token into the tokenizer vocabulary. To preserve spatial awareness and allow the model to distinguish among different viewpoints, we explicitly annotate each token with its corresponding camera

perspective in the prompt (e.g., CAM_FRONT: <image>, ..., CAM_BACK: <image>). These <image> tokens are dynamically substituted with the corresponding visual embeddings processed by the image encoder, enabling seamless integration of textual and visual modalities. To better support the implementation of NSP and DeCoT, we introduce six additional special tokens beyond <image>. The input and target sequences for the LLM are structured as follows:

User: {velocity} {acceleration} {navigation command} {image tokens}_t {prompt}.

Assistant: [BOS] [BOI] {image tokens}_t+3 [EOI] [BOT] {reasoning steps} [EOT] {generated trajectory} [EOS].

Where [BOS] and [EOS] are the original special tokens in the text tokenizer, [BOI] and [EOI] marking the start and end of the image tokens. Similarly, [BOT] and [EOT] represent the begin and end of reasoning process.

To perform human-like reasoning, we introduce explicit supervision over the language model's intermediate reasoning steps, thereby enhancing its ability to handle complex decision-making tasks. Leveraging both visual and textual modalities, ReasonPlan performs a planning-oriented thinking process before the final planning, including *Scene Understanding*, *Traffic Sign Recognition*, *Critical Object Identification for Risk Assessment*, and *Meta Action*. For a sequence of length L, we compute the probability of the target answers $\mathbf{X}_a$ and the Cross-Entropy (CE) loss:

$$p(\mathbf{X}_a|\mathbf{X}_v, \mathbf{X}_p) = \prod_{i=1}^{L} p(\boldsymbol{x}_i|\mathbf{X}_v, \mathbf{X}_{p,<i}, \mathbf{X}_{a,<i}), \mathcal{L}_{\text{text}} = -\log p(\mathbf{X}_a|\mathbf{X}_v, \mathbf{X}_p), \tag{4}$$

where $\mathbf{X}_{p,<i}$ and $\mathbf{X}_{a,<i}$ are the instruction and answer tokens in all turns before the current prediction token $\boldsymbol{x}_i$, respectively.

Unlike traditional multi-turn QAs paradigms, ReasonPlan performs multi-step reasoning in a single forward pass, effectively leveraging the commonsense reasoning capabilities of LLM.

3.3 Training Strategy

The final training objective is formulated as a weighted combination of the self-supervised image prediction loss and the language reasoning loss:

$$\mathcal{L}_{\text{total}} = \lambda_1 \cdot \mathcal{L}_{\text{image}} + \lambda_2 \cdot \mathcal{L}_{\text{text}} \tag{5}$$

where λ_1 and λ_2 are the weights for the visual-space MSE loss and the textual-space CE loss, respectively. The overall framework is optimized in two stages, as shown in Fig. 2 (c). In the first stage, we train the projection module and context encoder using non-decision supervision to align the visual feature space with the textual semantic space, while extracting contextual cues from ego vehicle states and navigation commands. In the second stage, we employ the collected PDM dataset to jointly fine-tune the projection module, context encoder, and the LLM backbone, transferring pretrained commonsense reasoning capabilities to complex driving scenarios.

4 PDR: Planning-oriented Decision Reasoning Dataset

While various QA datasets have been introduced for autonomous driving [13, 19, 20], they are primarily designed for scene understanding and open-loop evaluation, and thus fail to assess actual driving performance in closed-loop settings. To bridge this gap, we construct a large-scale, high-quality decision reasoning dataset, namely PDR, which focuses on trajectory planning and comprises 210k diverse and high-quality samples. We develop an automated annotation pipeline tailored for complex decision-making in closed-loop scenarios. PDR intends to leverage the LLM's reasoning and generalization capabilities in dynamic driving environments.

To construct a reliable reasoning dataset, we extend the ground-truth annotations provided by Bench2Drive [27] with structured reasoning labels. The reasoning traces have undergone thorough human verification to ensure consistency and interpretability across diverse scenarios. As shown

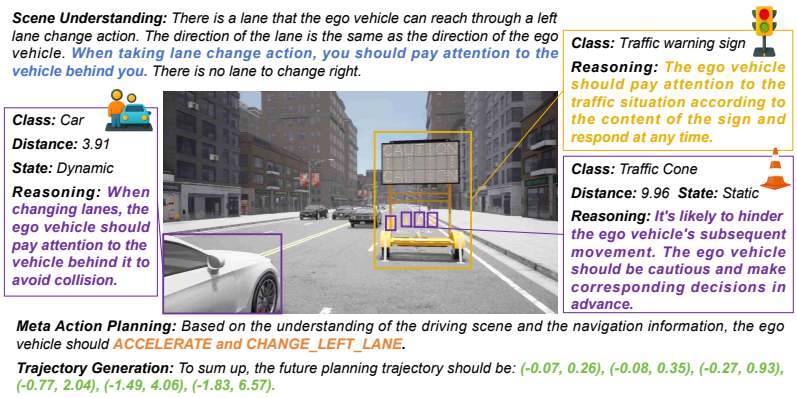


Figure 4: An annotated sample of the PDR dataset.

Table 1: **Planning and Multi-Ability Performance in Bench2Drive.** * denotes expert feature distillation. M=Merging. O=Overtaking. EB=Emergency Brake. Give Way=GW. TS=Traffic Sign.

Method	Open-loop	Closed-loop Metric				Ability (%) ↑					Ability
	Avg. L2 ↓	DS ↑	SR ↑	Effi ↑	Comf ↑	M	O	EB	GW	TS	Mean
TCP* [33]	1.70	40.70	15.00	54.26	47.80	17.50	13.63	20.00	10.00	6.81	13.59
ThinkTwice* [46]	0.95	62.44	31.23	69.33	16.22	13.72	22.93	52.99	50.00	47.78	37.48
DriveAdapter* [47]	1.01	**64.22**	**33.08**	70.22	16.01	14.55	22.61	54.04	50.00	50.45	**38.33**
AD-MLP [29]	3.64	18.05	0.00	48.45	22.63	0.00	0.00	0.00	0.00	0.00	0.00
UniAD-Tiny [1]	0.80	40.73	13.18	123.92	47.04	7.04	10.00	21.82	20.00	14.61	14.69
UniAD-Base [1]	0.73	45.81	16.36	129.21	43.58	12.16	20.00	23.64	10.00	13.89	15.94
VAD [2]	0.91	42.35	15.00	157.94	46.01	7.14	20.00	16.36	20.00	20.22	16.75
MomAD [24]	0.82	47.91	18.11	174.91	**51.20**	-	-	-	-	-	-
ReasonPlan (**Ours**)	**0.61**	**64.01**	**34.55**	**180.64**	25.63	**37.50**	**26.67**	**33.30**	**40.00**	**45.79**	**36.66**

in Fig. 4, the reasoning process comprises the following stages: *Scene Understanding*, *Traffic Sign Recognition*, *Critical Object Identification for Risk Assessment*, and *Meta Action*. In Appendix A, we detail the statistics of PDR and the annotation pipeline for each component.

5 Experiments

5.1 Benchmark and Evaluation Metrics

Benchmark. We evaluate the closed-loop driving performance of ReasonPlan on Bench2Drive [27], which features challenging interactive scenarios based on the Carla leaderboard v2. To further assess its reasoning capabilities, we also conduct zero-shot evaluations on DriveOcclusionSim (DOS) [28], a suite of complex scenarios requiring models to infer global context from dynamic driving environments. For ablation studies, we use the Dev10 [27] for quick validation.

Implementation details. Our framework processes six surround-view images captured at an original resolution 1600×900, which are resized and encoded using the AnyRes [44] strategy to produce ten spatial grids of size 384×384. Our framework employs SigLIP [45] as the vision encoder. For the language model, we adopt Qwen-0.5B [10], a lightweight yet capable LLM that balances efficiency and reasoning capability. The learning rate is fixed at 5e-5, and the weights for both $\mathcal{L}_{\text{image}}$ and $\mathcal{L}_{\text{text}}$ are set to 1.0. More details in Appendix C.

Evaluation Metrics. For open-loop, we report the L2 distance between 2s predicted trajectories and expert trajectories. For closed-loop, we adopt the metrics from [27]: (1) Route Completion (RC): percentage of completed route; (2) Infraction Score (IS): traffic violation penalty score; (3) Driving Score (DS): RC × IS (overall performance metric); (4) Success Rate (SR): percentage of infraction-free, timely completed episodes; (5) Efficiency (Effi): ego speed relative to neighboring vehicles' average; (6) Comfort (Comf): compliance with motion smoothness thresholds.

Table 2: **Planning Performance in DOS**. DOS includes four types of challenging occlusion-reasoning driving scenarios. DOS_01 = Parked Cars, DOS_02 = Sudden Brake, DOS_03 = Left Turn, DOS_04 = Red Light Infraction. The details about each scenario can be found in Appendix B.

Method	DOS_01			DOS_02			DOS_03			DOS_04			Average↑
	RC↑	IS↑	DS↑	RC↑	IS↑	DS↑	RC↑	IS↑	DS↑	RC↑	IS ↑	DS↑	
UniAD-Tiny [1]	85.98	0.58	49.95	90.91	0.78	69.74	50.01	**0.97**	49.83	89.25	0.687	62.18	57.93
UniAD-Base [1]	95.65	0.69	66.00	95.46	0.73	69.32	93.78	0.78	73.18	79.07	0.97	75.87	71.09
VAD [2]	86.97	0.71	58.46	78.80	0.78	57.84	74.75	0.85	61.16	82.91	0.82	65.88	60.84
MomAD [24]	64.86	**0.86**	53.90	82.71	0.80	63.36	64.87	0.91	56.00	55.06	**1.00**	55.06	57.08
LMDrive [18]	**100.00**	0.64	64.00	**100.00**	0.62	62.00	94.42	0.78	73.65	**91.16**	0.64	58.48	64.53
ReasonPlan (**Ours**)	99.78	0.68	**67.77**	99.65	**0.92**	**91.65**	**95.53**	0.84	**80.57**	90.65	0.87	**77.85**	**78.02**

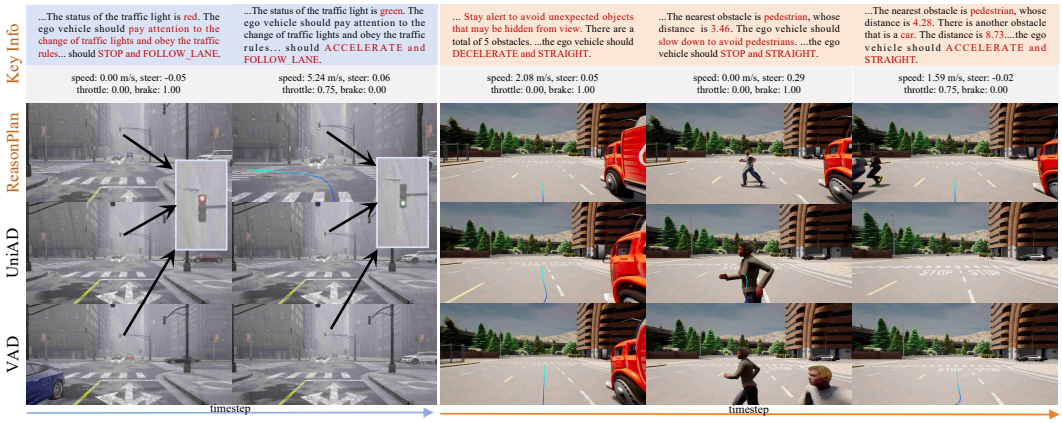


Figure 5: Qualitative comparison of ReasonPlan with baselines. The left case is the signalized junction within Bench2Drive. While baseline methods stall at green lights due to misinterpreting signal changes, ReasonPlan accurately detects the transition and proceeds safely through intersections. The right case is the pedestrian emerging scenario within DOS. While other methods fail to react in time and result in a collision, ReasonPlan anticipates the risk by decelerating early and executing a timely stop upon detection.

5.2 Comparison with State-of-the-arts

SOTA Performance in both Open-Loop and Closed-Loop Evaluations. As reported in Tab. 1, **ReasonPlan** achieves the best open-loop performance with **0.61** L2 error, demonstrating superior trajectory prediction accuracy. In closed-loop evaluation, ReasonPlan delivers competitive performance compared to SOTA method DriveAdapter [47] which utilizes privileged expert feature distillation. Specifically, it achieves a DS of **64.01**, significantly outperforming the non-distillation IL-based SOTA method, MomAD [24] by **16.1**. Furthermore, ReasonPlan improves the success rate by **16.44%** compared to MomAD [24]. ReasonPlan also achieves the **highest efficiency score of 180.64** among all evaluated methods, demonstrating an effective and proactive driving policy. The comfort score of 25.63 reflects a common trade-off between trajectory agility and smoothness. Nonetheless, the comfort margin could be further optimized via post-smoothing or low-level controller tuning. Besides, **ReasonPlan** achieves a **Mean Ability** score of **36.66%** across diverse driving scenarios, significantly outperforming all E2E baselines. These results underscore the model's strong capability to reason across diverse driving intents and validate its robustness under complex, multi-intent scenarios.

Strong Zero-Shot Generalization in Complex Reasoning Scenarios. To evaluate the out-of-distribution generalization capability of **ReasonPlan**, we conduct zero-shot closed-loop evaluations on the **DOS** benchmark, as summarized in Tab. 2. Notably, no methods are trained on DOS, ensuring a purely zero-shot setting. ReasonPlan demonstrates consistent and superior performance under these conditions, achieving the highest average DS of **78.02**. These results highlight the strong generalization capacity of ReasonPlan, driven by its holistic reasoning pipeline, which enables robust and safe decision-making even in unseen scenarios.

5.3 Qualitative Results

Fig. 5 illustrates qualitative results of ReasonPlan in two representative closed-loop evaluation scenarios. The figures showcase the DeCoT reasoning process and the corresponding predicted trajectories. Compared to baseline methods, ReasonPlan exhibits superior performance in navigating complex intersections and handling unseen scenarios.

5.4 Ablation Studies and Analysis

In this section, through detailed ablation studies, we validate the effectiveness of our proposed methods and datasets through model and reasoning steps ablations.

NSP effectively models dynamic scene transitions and enhances spatial planning. The NSP module introduces fine-grained visual understanding by applying dense supervision on image tokens. Through the temporally predictive task, it facilitates improved 3D spatial reasoning and enhances downstream planning performance (Tab. 3a ID 2). However, without explicit reasoning over decision-making, NSP alone fails to ensure traffic compliance, resulting in lower IS.

DeCoT consistently facilitates planning through structured reasoning. DeCoT decomposes complicated decision-making into interpretable reasoning steps with direct supervision, thereby improving the model's capacity to handle intricate scenarios (Tab. 3a ID 3).

NSP and DeCoT are complementary and synergistic . Integrating both NSP and DeCoT achieves the best overall performance (Tab. 3a ID 4), as NSP provides rich visual context for planning, while DeCoT enforces structured reasoning to regulate driving behavior. These components empower ReasonPlan to perform unified, interpretable, and effective E2E planning in complex scenarios.

Table 3: Ablations on (a) Each Module and (b) Reasoning Steps (Dev10). NSP=Next Scene Prediction. DeCoT=Decision Chain-of Thought. SU=Scene Understanding, TS=Traffic Sign Recognition, CO=Critical Object Identification for Risk Assessment, MA=Meta Action.

ID	NSP	DeCoT	RC↑	IS↑	DS↑
1	✗	✗	58.79	**0.79**	41.84
2	✓	✗	**95.03**	0.54	52.61
3	✗	✓	73.73	0.76	53.97
4	✓	✓	89.29	0.65	**57.83**

(a) Model architecture designs.

ID	SU	TS	CO	MA	RC↑	IS↑	DS↑
1	✗	✓	✓	✓	69.02	0.65	43.28
2	✓	✗	✓	✓	74.94	0.59	45.65
3	✓	✓	✗	✓	76.54	0.55	43.98
4	✓	✓	✓	✗	64.25	**0.73**	42.61
5	✓	✓	✓	✓	**89.29**	0.65	**57.83**

(b) Individual Reasoning Steps.

Structured and complete reasoning steps lead to the best driving performance. As shown in Tab. 3b, removing any single component from the full reasoning pipeline degrades the model's ability to handle complex scenarios. Specifically, omitting the *Meta Action* step, which represents the driving decision, leads to a significant drop in DS. These results emphasize the importance of fine-grained reasoning for safe and robust decision-making. Moreover, these findings underscore the quality and effectiveness of the PDR dataset generated by our automated annotation pipeline.

6 Conclusion

In this work, we present ReasonPlan, a novel fine-tuning framework that adapts MLLMs for complex closed-loop scenarios. ReasonPlan introduces a self-supervised Next Scene Prediction task and explicit Decision Chain-of-Thought process, enabling unified integration of visual and textual modalities for interpretable planning. Through comprehensive evaluations on Bench2Drive and DOS, ReasonPlan achieves excellent performance in both open-loop and closed-loop settings. Notably, it demonstrates strong zero-shot generalization on unseen tasks, underscoring its robustness and potential for real-world deployment. Our results highlight the promise of MLLM-based frameworks in bridging high-level reasoning and low-level planning, paving the way for more cognitive and generalizable autonomous driving systems.

7 Limitations

Firstly, while ReasonPlan leverages strong reasoning capabilities to address challenges in complex reasoning and out-of-distribution scenarios, its reliance on MLLMs introduces non-negligible inference latency. Although the LLM of 0.5B size is currently acceptable on real-time deployment, larger sizes such as 7B are still difficult to deploy in real time. Nonetheless, emerging techniques in latent-space reasoning offer promising directions to improve both the efficiency and responsiveness of the framework. In addition, can large reasoning models such as o1 or DeepSeek-R1 provide better reasoning capabilities for closed-loop autonomous driving is worth further analysis. Secondly, ReasonPlan employs a holistic reasoning framework where actions are represented as unimodal textual outputs. A promising future direction is to decouple reasoning and action generation—delegating decision-making solely to the LLM, while using a specialized generative model to synthesize multimodal trajectories. This modular design, inspired by recent successes in robotics, may further enhance flexibility and scalability. Thirdly, similar to most end-to-end frameworks, ReasonPlan relies on supervised fine-tuning over offline datasets, which limits its ability to learn from interactive feedback and may lead to occasional collisions (see Appendix E). Post-training with reinforcement learning or integrating environment-aware adaptation mechanisms might be a future direction to address this limitation. Last but not least, it remains unclear whether MLLMs are the most suitable foundation model for end-to-end autonomous driving and can well align visual language and actions. The VLA foundation model for autonomous driving should be further investigated.

8 Acknowledgement

This work was supported by the National Key Re-search and Development Program of China under Grant 2022YFA1004000, in part by Beijing Natural Science Foundation under Grant L253007 and 4242052 , in part by the State Key Laboratory of Safety Intelligent Mining in Non-coal Open-pit Mines, National Mine Safety Administration (Grant No.2024-ZD04).

References

[1] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, et al. Planning-oriented autonomous driving. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023.

[2] B. Jiang, S. Chen, Q. Xu, B. Liao, J. Chen, H. Zhou, Q. Zhang, W. Liu, C. Huang, and X. Wang. Vad: Vectorized scene representation for efficient autonomous driving. *Porc. of IEEE Intl. Conf. on Computer Vision*, 2023.

[3] P. Yang, Y. Zheng, Q. Zhang, K. Zhu, Z. Xing, Q. Lin, Y.-F. Liu, Z. Su, and D. Zhao. Uncad: Towards safe end-to-end autonomous driving via online map uncertainty. *Proc. of the IEEE Int. Conf. on Robotics and Automation*, 2025.

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

[5] W. Sun, X. Lin, Y. Shi, C. Zhang, H. Wu, and S. Zheng. Sparsedrive: End-to-end autonomous driving via sparse scene representation. *Proc. of the IEEE Int. Conf. on Robotics and Automation*, 2024.

[6] P. De Haan, D. Jayaraman, and S. Levine. Causal confusion in imitation learning. *Proc. of Advances in Neural Information Processing Systems*, 32, 2019.

[7] S. Ross, G. Gordon, and D. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proc. of Intl Conf. on Artificial Intelligence and Statistics*, pages 627–635, 2011.

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

[9] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

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

[11] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, H. Yang, et al. Deepseek-vl: Towards real-world vision-language understanding. *CoRR*, 2024.

[12] J. Chai, S. Li, Y. Fu, D. Zhao, and Y. Zhu. Empowering llm agents with zero-shot optimal decision-making through q-learning. In *Proc. of Intl. Conf. on Learning Representations*, 2025.

[13] W. Wang, J. Xie, C. Hu, H. Zou, J. Fan, W. Tong, Y. Wen, S. Wu, H. Deng, Z. Li, et al. Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *arXiv preprint arXiv:2312.09245*, 2023.

[14] X. Tian, J. Gu, B. Li, Y. Liu, Y. Wang, Z. Zhao, K. Zhan, P. Jia, X. Lang, and H. Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. In *Conf. on Robot Learning*, 2024.

[15] B. Jiang, S. Chen, B. Liao, X. Zhang, W. Yin, Q. Zhang, C. Huang, W. Liu, and X. Wang. Senna: Bridging large vision-language models and end-to-end autonomous driving. *arXiv preprint arXiv:2410.22313*, 2024.

[16] C. Pan, B. Yaman, T. Nesti, A. Mallik, A. G. Allievi, S. Velipasalar, and L. Ren. Vlp: Vision language planning for autonomous driving. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 14760–14769, 2024.

[17] D. Hegde, R. Yasarla, H. Cai, S. Han, A. Bhattacharyya, S. Mahajan, L. Liu, R. Garrepalli, V. M. Patel, and F. Porikli. Distilling multi-modal large language models for autonomous driving. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2025.

[18] H. Shao, Y. Hu, L. Wang, G. Song, S. L. Waslander, Y. Liu, and H. Li. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 15120–15130, 2024.

[19] T. Tian, B. Li, X. Weng, Y. Chen, E. Schmerling, Y. Wang, B. Ivanovic, and M. Pavone. Tokenize the world into object-level knowledge to address long-tail events in autonomous driving. In *Conf. on Robot Learning*, 2024.

[20] M. Nie, R. Peng, C. Wang, X. Cai, J. Han, H. Xu, and L. Zhang. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. In *Proc. of European Conference on Computer Vision*, pages 292–308. Springer, 2024.

[21] L. Wen, D. Fu, X. Li, X. Cai, M. Tao, P. Cai, M. Dou, B. Shi, L. He, and Y. Qiao. Dilu: A knowledge-driven approach to autonomous driving with large language models. In *Proc. of Intl. Conf. on Learning Representations*, 2024.

[22] Z. Yang, X. Jia, H. Li, and J. Yan. Llm4drive: A survey of large language models for autonomous driving. In *Advances in Neural Information Processing Systems 2024 Workshop on Open-World Agents*.

[23] W. Zheng, Z. Xia, Y. Huang, S. Zuo, J. Zhou, and J. Lu. Doe-1: Closed-loop autonomous driving with large world model. *arXiv preprint arXiv:2412.09627*, 2024.

[24] Z. Song, C. Jia, L. Liu, H. Pan, Y. Zhang, J. Wang, X. Zhang, S. Xu, L. Yang, and Y. Luo. Don't shake the wheel: Momentum-aware planning in end-to-end autonomous driving. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2025.

[25] Y. Gao, Q. Zhang, D.-W. Ding, and D. Zhao. Dream to drive with predictive individual world model. *IEEE Transactions on Intelligent Vehicles*, pages 1–16, 2024. doi:10.1109/TIV.2024.3408830.

[26] Y. Li, L. Fan, J. He, Y. Wang, Y. Chen, Z. Zhang, and T. Tan. Enhancing end-to-end autonomous driving with latent world model. In *Proc. of Intl. Conf. on Learning Representations*, 2025.

[27] X. Jia, Z. Yang, Q. Li, Z. Zhang, and J. Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. *Proc. of Advances in Neural Information Processing Systems*, 2024.

[28] H. Shao, L. Wang, R. Chen, S. L. Waslander, H. Li, and Y. Liu. Reasonnet: End-to-end driving with temporal and global reasoning. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2023.

[29] J.-T. Zhai, Z. Feng, J. Du, Y. Mao, J.-J. Liu, Z. Tan, Y. Zhang, X. Ye, and J. Wang. Rethinking the open-loop evaluation of end-to-end autonomous driving in nuscenes. *arXiv preprint arXiv:2305.10430*, 2023.

[30] B. Liao, S. Chen, H. Yin, B. Jiang, C. Wang, S. Yan, X. Zhang, X. Li, Y. Zhang, Q. Zhang, et al. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2025.

[31] Z. Xing, X. Zhang, Y. Hu, B. Jiang, T. He, Q. Zhang, X. Long, and W. Yin. Goalflow: Goal-driven flow matching for multimodal trajectories generation in end-to-end autonomous driving. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2025.

[32] D. Dauner, M. Hallgarten, T. Li, X. Weng, Z. Huang, Z. Yang, H. Li, I. Gilitschenski, B. Ivanovic, M. Pavone, et al. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *Advances in Neural Information Processing Systems*, 37:28706–28719, 2024.

[33] P. Wu, X. Jia, L. Chen, J. Yan, H. Li, and Y. Qiao. Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. In *Proc. of Advances in Neural Information Processing Systems*, 2022.

[34] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. Carla: An open urban driving simulator. In *Conf. on Robot Learning*, pages 1–16, 2017.

[35] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Proc. of Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[36] H. Shao, S. Qian, H. Xiao, G. Song, Z. Zong, L. Wang, Y. Liu, and H. Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Proc. of Advances in Neural Information Processing Systems*, 37: 8612–8642, 2024.

[37] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, P. R. Sanketi, Q. Vuong, et al. Openvla: An open-source vision-language-action model. In *Conf. on Robot Learning*, 2024.

[38] S. Tu, J. Sun, Q. Zhang, X. Lan, and D. Zhao. Online preference-based reinforcement learning with self-augmented feedback from large language model. *In Proc. of Intl. Conf. on Autonomous Agents and Multiagent Systems*, 2024.

[39] Y. Chen, S. Tian, S. Liu, Y. Zhou, H. Li, and D. Zhao. Conrft: A reinforced fine-tuning method for vla models via consistency policy. *In Proc. of Robotics: Science and Systems*, 2025.

[40] A. Prakash, K. Chitta, and A. Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 7077–7087, 2021.

[41] E. Leurent. An environment for autonomous driving decision-making. https://github.com/eleurent/highway-env, 2018.

[42] K. Renz, L. Chen, E. Arani, and O. Sinavski. Simlingo: Vision-only closed-loop autonomous driving with language-action alignment. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2025.

[43] K. Renz, L. Chen, A.-M. Marcu, J. Hünermann, B. Hanotte, A. Karnsund, J. Shotton, E. Arani, and O. Sinavski. Carllava: Vision language models for camera-only closed-loop driving. *arXiv preprint arXiv:2406.10165*, 2024.

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

[45] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. In *Porc. of IEEE Intl. Conf. on Computer Vision*, pages 11975–11986, 2023.

[46] X. Jia, P. Wu, L. Chen, J. Xie, C. He, J. Yan, and H. Li. Think twice before driving: Towards scalable decoders for end-to-end autonomous driving. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2023.

[47] X. Jia, Y. Gao, L. Chen, J. Yan, P. L. Liu, and H. Li. Driveadapter: Breaking the coupling barrier of perception and planning in end-to-end autonomous driving. In *Porc. of IEEE Intl. Conf. on Computer Vision*, 2023.

ReasonPlan: Unified Scene Prediction and Decision Reasoning for Closed-loop Autonomous Driving

Supplementary Material

We provide supplementary material to complement the main paper, arranged as follows:

- Appendix A: Details on the PDR dataset.
- Appendix B: Benchmarks.
- Appendix C: Implementation Details.
- Appendix D: More Ablation Experiments.
- Appendix E: More Case Study.

A Details on the PDR dataset

A.1 Annotation Procedure

Our dataset is sourced from Bench2Drive [27], an offline dataset collected in highly challenging scenarios from Carla Leaderboard v2. Bench2Drive leverages the CARLA simulator to collect primitive per-frame annotations, including ego-vehicle states, sensor data, and detailed metadata on surrounding vehicles, pedestrians, and traffic signs. This enables a fully automated annotation process, eliminating the potential subjectivity and labor-intensiveness of manual labeling. Furthermore, it provides accurate information such as the relative distances between the ego vehicle and surrounding entities, significantly enhancing the dataset's accuracy and reliability for reasoning tasks. Specifically, to ensure high-quality reasoning data, the entire annotation process consists of the following stages:

***Scene Understanding*.** In autonomous driving, the ego vehicle must continuously perceive and reason about its surroundings to enable safe and efficient trajectory planning. Leveraging high-definition (HD) map information, our system accurately identifies the driving context—such as highways, intersections, and ramp entrances—by parsing road topology and lane semantics, including lane count, types, and connectivity. Based on the recognized scenario, the system delineates the ego vehicle's drivable area and enumerates all feasible candidate routes. These include adjacent lanes for potential lane changes, upcoming forks, and reachable target lanes. Such structured scene understanding serves as a critical foundation for downstream decision-making and motion planning, enabling the formulation of informed lane-changing and turning strategies.

***Traffic Sign Recognition*.** Our dataset includes comprehensive annotations for diverse traffic signage, encompassing traffic signals, warning signs, and construction indicators. Warning and construction signs are particularly essential, as they offer prior knowledge about potential hazards such as congestion or partial lane closures. Upon detection, the system anticipates disruptions and adapts its plan through maneuvers such as preemptive lane changes. For traffic lights, we explicitly annotate three discrete signal states—red, green, and yellow—allowing for fine-grained decision logic in regulated environments.

***Critical Object Identification for Risk Assessment*.** Robust identification of critical dynamic objects and the assessment of their associated risk are fundamental to safe motion planning. We annotate all potentially influential obstacles within each scene, filtering candidates based on interactive lane semantics and proximity to the ego vehicle. Using ground-truth 3D annotations, we extract fine-grained object attributes, including category, position, and motion state. These features are further used to reason about interaction dynamics—such as whether an object may trigger acceleration, deceleration, lane changes, or emergency maneuvers. The resulting risk estimations are translated into textual representations, facilitating downstream planning modules grounded in language-based reasoning.

Instruction prompt

You are a multimodal autonomous driving model. The ego car has a total of six cameras: front view, left front, right front, back view, left back and right back. The pictures they take are CAM_FRONT: <image>, CAM_FRONT_LEFT: <image>, CAM_FRONT_RIGHT: <image>, CAM_BACK: <image>, CAM_BACK_LEFT: <image>, CAM_BACK_RIGHT: <image>. The current state of the ego vehicle is described as follows: the current vehicle's longitudinal speed is 7.72 m/s, and the longitudinal acceleration is -0.96 m/s^2. The history trajectory is (0.00, -8.36), (0.00, -8.21), (0.00, -6.83), (-0.00, -3.93), (0.00, 0.00) and the current target point is (-0.05, 44.01). Based on the multi-view camera images of the ego vehicle and the above information, predict a future trajectory over the next 3 seconds, including 6 waypoints.

Answer

<think>*There is a lane that the ego vehicle can reach through a left lane change action. The direction of the lane is opposite to the direction of the ego vehicle. When taking lane change action, you should pay attention to the direction of the traffic. There is no lane to change right. In the current lane of the ego vehicle, there are a total of 3 obstacles ahead. There's a traffic sign ahead, and it's accident warning sign. The ego vehicle should pay attention to the traffic situation according to the content of the sign and respond at any time. The information about the three obstacles closest to the ego vehicle is as follows. The nearest obstacle to the ego vehicle is car, whose distance from the ego vehicle is 9.39. Its motion state is dynamic. In addition, there is another obstacle that is a car. The distance between the obstacle and ego vehicle is 27.94. Its motion state is dynamic. There are no obstacles ahead in the left lane adjacent to the ego vehicle's lane. Based on the understanding of the driving scene and the navigation information, the ego vehicle should KEEP_SPEED and FOLLOW_LANE.***</think>** *To sum up, the future planning trajectory should be: (0.00, 4.02), (0.00, 7.99), (0.00, 11.87), (-0.00, 15.89), (-0.00, 19.52), (-0.00, 23.30).*

Figure A1: An Complete sample of the PDR dataset.

***Meta Action*.** To emulate high-level human driving intent, we annotate expert trajectories with structured meta-actions across lateral and longitudinal dimensions. Lateral decisions include six motion primitives: lane follow, left/right lane change, left/right turn, and straight, capturing diverse driving intents across typical scenarios. For longitudinal behavior, we annotate acceleration, deceleration, stop, and emergency braking to reflect varying control responses. Lateral labels are derived from observed deviations in trajectory geometry, while longitudinal annotations follow expert speed profiles. Given the noise inherent in reinforcement learning (RL)-based expert policies, we apply a low-pass filter to suppress high-frequency fluctuations and enhance annotation stability. Finally, these meta actions are serialized into language-based formats to support integration with text-conditioned planning modules.

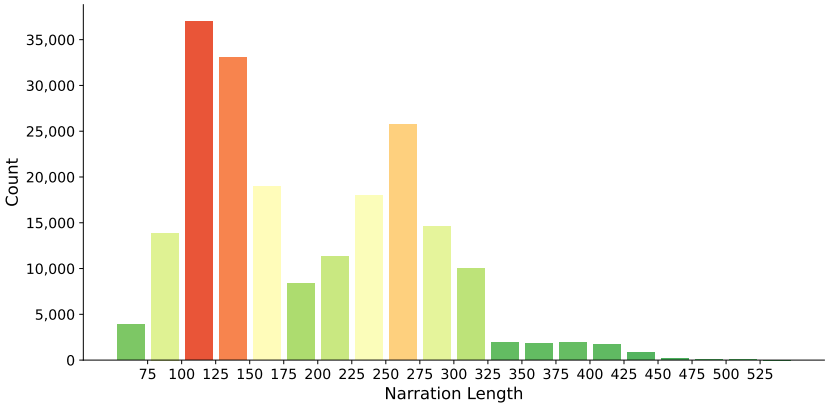
A.2 Data Statistics

Through our automated pipeline, we constructed a high-quality reasoning dataset containing 203,353 annotated instances, with an average of 195 words per image frame. Fig. A1 shows an Complete sample of the PDR dataset. The dataset's attributes are illustrated in Fig. A2. The narration length is up to 500 words, while the vocabulary chart indicates the planning-oriented information.

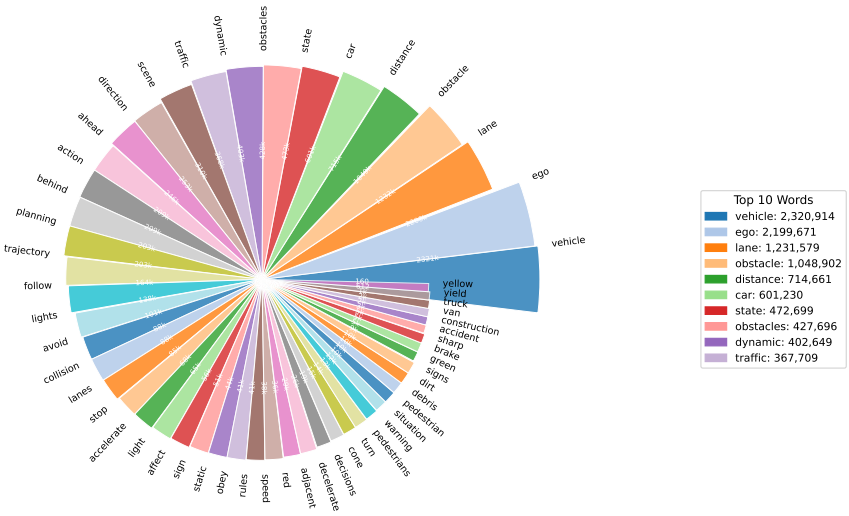
B Benchmarks

B.1 Bench2Drive

Bench2Drive [27] designs a comprehensive and fair testing benchmark for evaluating E2E autonomous driving systems' multiple abilities in a closed-loop manner. Bench2Drive is an offline



(a)



(b)

Figure A2: Data statistics of PDR. (a) Distribution of reasoning sentence length. (b) The key vocabulary chart of PDR.

dataset collected using the RL-based agent in CARLA v2. We adopt the Base subset, which contains 230k frames extracted from 1,000 short clips. After filtering out the terminal frames of each scenario, a total of 203,353 samples are retained. These clips are uniformly distributed across 44 interactive driving scenarios, 23 weather conditions, and 12 towns. The evaluation protocol requires E2E autonomous driving models to complete all 44 scenarios under various locations and weather conditions, resulting in 220 distinct routes. This setup enables a comprehensive and disentangled assessment of driving capabilities across diverse conditions.

Due to the high computational cost of full evaluation, we employ the Dev10 subset for ablation studies. It consists of 10 carefully selected routes from the official set, chosen to be both challenging and representative while maintaining low variance.

B.2 DOS

DOS is publicly released in ReasonNet [28], a driving simulation benchmark in Carla consisting of diverse occlusion events, where E2E models have to reason about the global information of driving scenes. DOS includes four types of challenging scenarios:

- **Parked Cars (DOS01):** The ego vehicle drives along a straight lane with parked vehicles on both sides. Pedestrians initially appear on the visible sidewalk but may suddenly emerge from occluded regions between parked cars, creating abrupt crossing hazards.
- **Sudden Brake (DOS02):** The ego vehicle follows traffic in a straight lane. Pedestrians unexpectedly emerge from the sidewalk, prompting leading vehicles to brake abruptly. These pedestrians remain outside the ego vehicle's field of view, making anticipation challenging.
- **Left Turn (DOS03):** The ego vehicle attempts an unprotected left turn at an intersection. A large truck in the opposing lane intermittently occludes oncoming traffic, making it difficult to detect vehicles proceeding straight.
- **Red Light Infraction (DOS04):** The ego vehicle crosses an intersection behind several trucks. A vehicle running a red light approaches laterally but remains occluded by the trucks until the last moment, leading to a high-risk collision scenario.

Each of the four scenarios in the DOS benchmark consists of 25 distinct cases with diverse road layouts and background traffic conditions, resulting in a total of 100 evaluation cases.

Table A1: Other parameters.

Parameters	Description	Value
L_v	the number of visual tokens per grid	729
D_v	the visual embedding dimension	1152
D_p	the textual embedding dimension	1024
b_s	the training batch size	16
l_r	the learning rate	5e-5

C Implementation Details

All experiments are conducted on 8 NVIDIA L20 GPUs (48 GB each). To accelerate the alignment of the vision-language space and gradually enhance the reasoning and planning capabilities of ReasonPlan, we adopt a two-stage training strategy. In each stage, the model inherits the weights from the previous stage and continues training. The first stage consists of 1 epoch and completes in approximately 6 hours, while the second stage spans 3 epochs, requiring around 23 hours. The other parameters are as shown in Tab. A1.

Table A2: Ablation on NSP Views.

Views	RC↑	IS↑	DS↑
Front-view	**89.29**	0.65	57.83
All-views	87.35	**0.66**	**58.95**

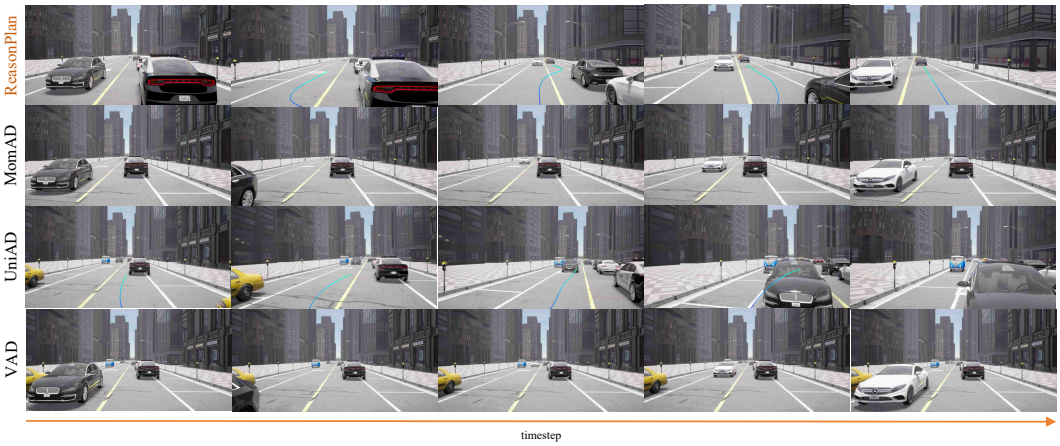
Table A3: Ablation on Image Weights

Image Weights	RC↑	IS↑	DS↑
0	73.73	**0.76**	53.97
0.25	80.12	0.68	56.96
0.5	79.52	0.75	**61.13**
1	**89.29**	0.65	57.83
2	80.80	0.66	54.44

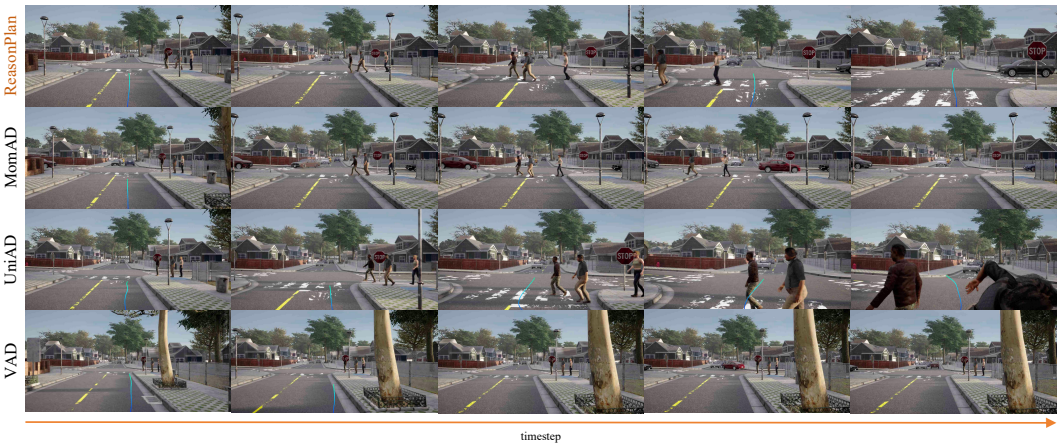
D More Ablation Experiments

Front-view prediction performs on par with using all views. For the next scene prediction (NSP) task, we conduct an ablation comparing models trained with only the front-view image versus those utilizing all camera views. As shown in Tab. A2, the results indicate that front-view-only prediction yields performance comparable to using the full view set. This suggests that the front-facing image captures the most semantically informative content for driving decisions, including cues critical for trajectory planning. To improve computational efficiency without compromising performance, we adopt the front view exclusively in all subsequent experiments.

A synergy–conflict dynamic exists between NSP and textual reasoning. As shown in Tab. A3, when the image loss weight is too low, the supervisory signal from NSP becomes negligible, offering



(a) AccidentTwoWays in Bench2Drive. The autonomous vehicle must execute two consecutive lane changes while maintaining awareness of oncoming traffic. MomAD and VAD, due to their conservative policies, come to a halt in front of the obstructing vehicle; UniAD, on the other hand, takes a more aggressive approach, disregarding the oncoming traffic, resulting in a collision during the lane change. In contrast, ReasonPlan accurately perceives the relative positions of surrounding vehicles, avoiding two successive oncoming vehicles in a timely manner and successfully completing the lane change task.



(b) PedestrianCrossing in Bench2Drive. After avoiding the pedestrian, both MomAD and VAD fail to resume their motion, while UniAD cannot react to the pedestrian's presence, leading to a collision. ReasonPlan, however, promptly halts upon detecting the pedestrian and successfully resumes operation once the pedestrian has cleared, completing the task efficient.

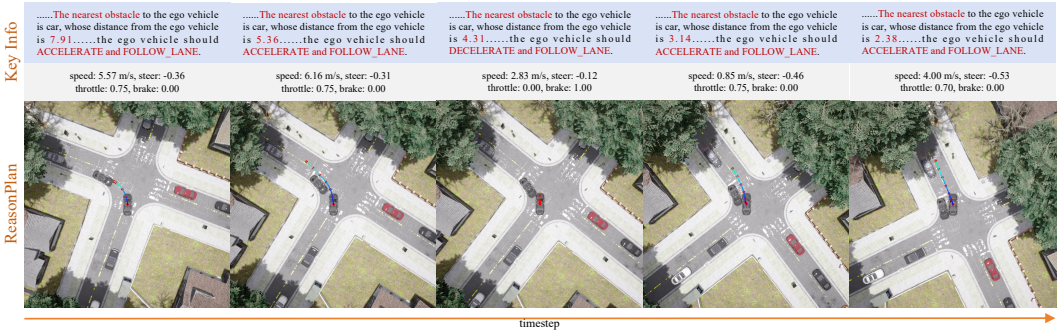
Figure A3: Additional Qualitative Comparison Cases for ReasonPlan.

limited benefit to language-based reasoning. In contrast, overly high image weights lead the model to overfit to visual prediction, impairing its reasoning capability in the textual domain. Empirically, setting the image weight to 0.5 provides a balanced trade-off, enabling effective interaction between NSP and DeCoT and yielding the best overall performance.

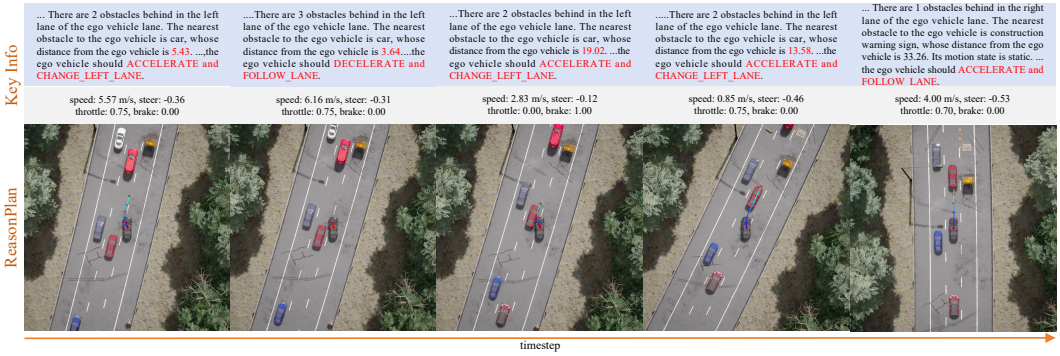
E More Case Study

We present two additional representative comparison cases in Fig. A3. In contrast, existing methods generally exhibit delayed responses or policy failures in such high-risk scenarios, further highlighting the advantages of ReasonPlan in terms of decision-making reliability and safety.

Although our fine-tuning framework enhances the model's reasoning capability in complex scenarios, failure modes still exist in some cases. As shown in FIg. A4, ReasonPlan still encounters unavoidable collision scenarios. As noted in the Limitations section, the method relies solely on



(a) Although identifying proximal obstacles, ReasonPlan exhibited failure in obstacle avoidance.



(b) ReasonPlan accurately detected the approaching vehicle from behind but failed to complete the lane change maneuver within a short time frame, resulting in a collision.

Figure A4: Failed Cases for ReasonPlan.

supervised fine-tuning with offline datasets and cannot incorporate feedback-driven learning. Reinforcement learning, as a trial-and-error learning paradigm, may present a promising direction to mitigate this constraint.