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

Structured Labeling Enables Faster Vision-Language Models for End-to-End Autonomous Driving

Hao Jiang[1], Chuan Hu[1,✉], Yukang Shi[2], Yuan He[2], Ke Wang[2], Xi Zhang[1], Zhipeng Zhang[1,✉]

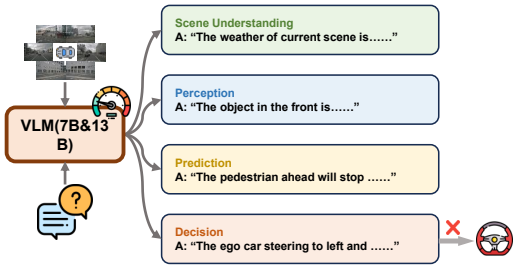
***Abstract*— Vision-Language Models (VLMs) offer a promising approach to end-to-end autonomous driving due to their human-like reasoning capabilities. However, troublesome gaps remains between current VLMs and real-world autonomous driving applications. One major limitation is that existing datasets with loosely formatted language descriptions are not machine-friendly and may introduce redundancy. Additionally, high computational cost and massive scale of VLMs hinder the inference speed and real-world deployment. To bridge the gap, this paper introduces a structured and concise benchmark dataset, NuScenes-S, which is derived from the NuScenes dataset and contains machine-friendly structured representations. Moreover, we present FastDrive, a compact VLM baseline with 0.9B parameters. In contrast to existing VLMs with over 7B parameters and unstructured language processing(e.g., LLaVA-1.5), FastDrive understands structured and concise descriptions and generates machine-friendly driving decisions with high efficiency. Extensive experiments show that FastDrive achieves competitive performance on structured dataset, with approximately 20% accuracy improvement on decision-making tasks, while surpassing massive parameter baseline in inference speed with over 10× speedup. Additionally, ablation studies further focus on the impact of scene annotations (e.g., weather, time of day) on decision-making tasks, demonstrating their importance on decision-making tasks in autonomous driving.**

I. Introduction

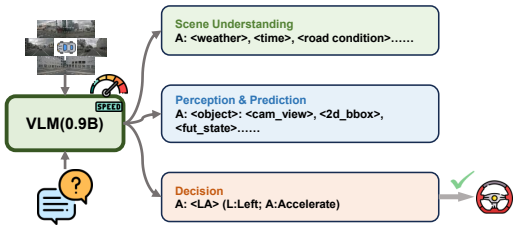
The rapid evolution of autonomous driving systems demands robust environmental understanding capabilities that transcend conventional perception modules [1], [2]. The integration of human-like reasoning into autonomous driving systems has become a pivotal research frontier, where Vision-Language Models (VLMs) have emerged as a transformative paradigm, offering human-like reasoning through multimodal fusion of visual inputs and linguistic context. While recent studies have shown the potential of VLMs in scene understanding and decision explanation [3], [4], [5], [6], critical gaps persist in real-world deployment: inefficient linguistic processing and computational overhead from model scale, which hinder performance and integration into autonomous driving systems.

Current VLMs training paradigms heavily rely on datasets with free-form textual annotations (Fig. 1), such as NuScenes-QA [7] and BDD-X [8]. While these descriptions capture rich semantic information, their syntactic variability forces models to parse redundant linguistic patterns. For example, the sentence “A black sedan is turning left” and “A sedan that is black is making a left turn” convey the same information but differ in structure, this syntactic variability increases the complexity of the learning task and computational overhead, as VLMs must disambiguate synonymous expressions rather than focus on core reasoning tasks. Additionally, another sentence ” Moving pedestrian wearing a white top and gray shorts in the crosswalk” along with the previous example contain redundant information, such as the color of the vehicle and the pedestrian’s clothing, which could introduce unnecessary cognitive load for downstream decision-making processes in autonomous driving systems. In this context, the VLMs may spend significant attention on irrelevant information rather than focusing on core event reasoning, causing wastage of computational resources and hindering inference efficiency. Moreover, we also observe that existing some baselines often rely on large-scale VLMs, such as DriveLM [4], DriveVLM [5], and LeapAD[6], etc, which tpically exceed 7B parameters or more to achieve multimodal alignment and reasoning. Although ultra-large parameter parameters VLMs may achieve fair performance in various benchmarks, along with the high computational cost, memory consumption, and inference latency, which rendering them impractical for real-time deployment in autonomous driving systems.

To address these challenges, this paper introduces a struc-



(a) Existing VLMs baseline



(b) FastDrive (Ours)

Fig. 1: Existing VLMs heavily rely on massive-parameter VLMs and free-form textual annotations, which introduce computational overhead and hinder inference efficiency. FastDrive, a compact VLM baseline for end-to-end autonomous driving with structured data, which enhances inference efficiency and integration into autonomous driving systems.

[1]Shanghai Jiao Tong University, [2]KargoBot, [✉]Corresponding author: {chuan.hu, zhipengzhang}@sjtu.edu.cn

tured and concise benchmark dataset, NuScenes-S, derived from the NuScenes dataset [9]. Different from existing datasets that feature free-form textual annotations with redundant information, NuScenes-S extract and summarize key elements that may affect driving decisions, such as vehicle type, vehicle action, pedestrian action, traffic light status, etc., into clear and concise phrases, and organize them into structured dictionarie format. By converting key information into structured key-value pairs, it ensures data consistency and significantly reduces the computational cost associated with natural language parsing. This structured representation allows for efficient retrieval of relevant information while filtering out redundant content, thereby enhancing the clarity and relevance of the input to downstream modules. Furthermore, it allows for the flexible construction of tailored question-answer pairs, which not only facilitates targeted model training but also provides a more effective and interpretable framework for comprehensive model evaluation. Additionally, a compact VLM baseline referenced from InternVL [10] is introduced, named FastDrive, which is designed for end-to-end autonomous driving with small-scale parameters. FastDrive mimics the reasoning strategies of human drivers by employing a chain-of-thought process to perform scene understanding, perception, prediction, and decision-making tasks, thereby achieving effective alignment with end-to-end autonomous driving frameworks. In summary, the main contributions of this paper are as follows:

- We introduce a structured dataset that focuses on key elements closely related to driving decisions, which eliminates redundant information and addresses the limitation of synonymous expressions in free-form textual annotations and enhances the efficiency of inference.
- A compact VLM baseline with 0.9B parameters is proposed, which mimics the reasoning strategies of human drivers and achieves effective alignment with end-to-end autonomous driving frameworks.
- A comprehensive evaluation and extensive experiments tailored for NuScenes-S and FastDrive are conducted. The results demonstrate the effectiveness of the proposed dataset and model, which achieves competitive performance on the NuScenes-S benchmark.

II. Related work

A. Driving with VLMs

VLMs [11], [12], [13] have emerged as a transformative paradigm in artificial intelligence, bridging multimodal understanding between visual inputs and linguistic context. These models have shown remarkable performance in various vision-language tasks, such as image captioning [14], visual question answering [15], [16], and visual reasoning [17]. These capabilities have enabled VLMs to understand complex visual scenes and generate plausible reasoning chains, making them a promising approach for end-to-end autonomous driving [18]. Early works [19], [20] have explored text dialogue capabilities of LLMs for autonomous planning tasks, but relying on handcrafted rules and linguistic description makes it difficult for LLMs fully understand driving scenarios. With the advent of VLMs, more recent works use VLMs to interact directly with driving environment through visual and linguistic inputs [21]. LMDriver [22] and DriveMLM [23] interact with dynamic driving environment through multimodal sensor inputs and natural language commands and directly output control commands, which treats VLMs as a black box and does not provide explicit reasoning process. DriveLM [4] devides the driving task into perception, prediction, and planning, and uses graph visual question answering to improve the interpretability of the reasoning process. DriveVLM [5] and Senna [24] adopts a novel approach by combining VLMs with modular end-to-end driving pipeline to compensate the shortcomings of modular black-box pipeline. LeapAD [6] inspired by human cognitive process, introduces three different VLMs to mimic scene understanding, decision making, and reflection process. Despite the progress, existing VLMs are not impractical for real-time inference or deployment in autonomous systems due to their large number of parameters and unstructured language descriptions in existing datasets.

B. Driving Datasets for VLMs

General-purpose VLMs struggle to achieve satisfactory performance in autonomous driving tasks due to challenges in dynamic scene understanding and multimodal reasoning. Traditional autonomous driving datasets, such as KITTI [25], Waymo Open Dataset [26], and NuScenes [27] mainly provide rich multimodal sensor data for perception or prediction tasks unsuitable for VLMs. In order to adapt to VLMs, Talk2Car [28], NuPrompt [29], NuScenes-QA [30] and DriveLM [4] introduce free-form language descriptions and QA pairs to the NuScenes dataset. BDD-X [31] and BDD-OIA [32] provide text annotations describing vehicle actions and their rationales. DRAMA [33] focus on driving hazards and related objects, this dataset provides rich visual scenes and object-level queries. Rank2Tell [34] annotates various semantic, spatial, temporal, and relational attributes of various important objects in complex traffic scenarios. Some of these datasets mainly focus on scene understanding, risk assessment, object-level queries, or multimodal reasoning, but lack structured language descriptions and reasoning chains, which could improve VLMs' capacity for understanding driving scenarios and enhance inference efficiency. Other datasets may include fairly complete autonomous driving tasks, but the language descriptions are unstructured and verbose, which hinders the integration of VLMs into autonomous systems. Our NuScenes-S manages dataset in a human-like reasoning manner across perception, prediction, and decision-making tasks, which focus on key elements in driving scenarios and convert unstructured language descriptions into structured and concise format, further improving efficiency and integration.

III. The NuScenes-Structured Benchmark

A. Scene Description

Understanding the driving scenario is crucial for making safe driving decisions. Therefore, the scene description in NuScenes-S is introduced to provide a more comprehensive

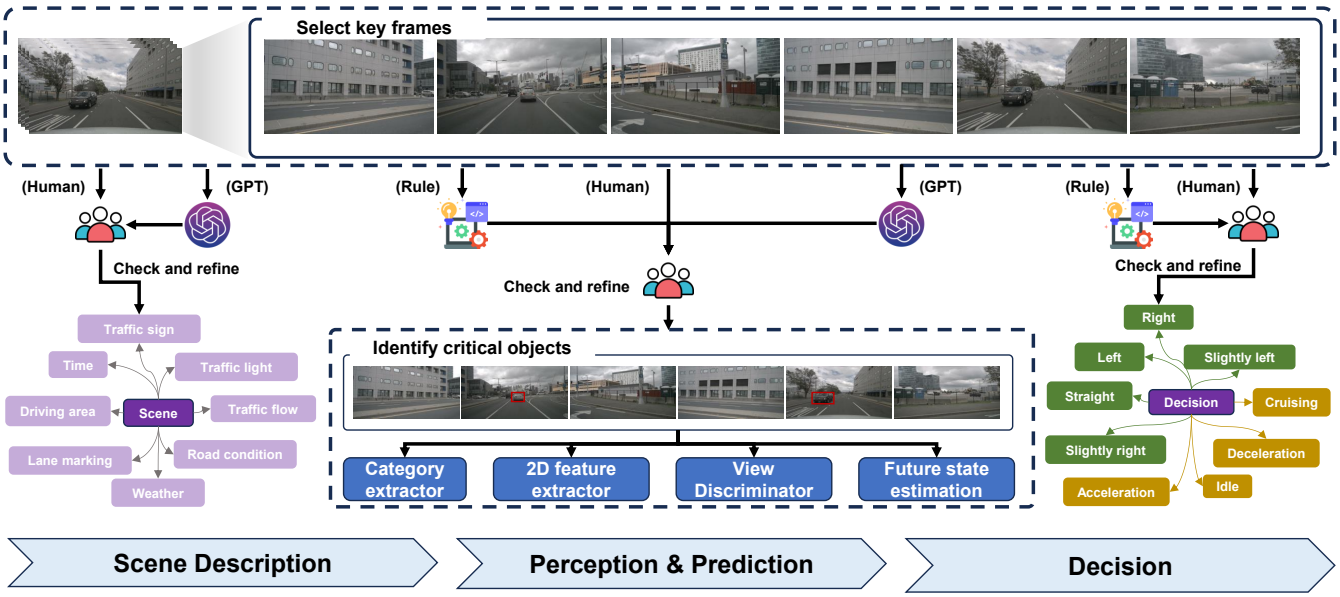


Fig. 2: The dataset construction process of the NuScenes-S dataset.

Dataset	Scene Description								Per.	Pre.	Dec.	Frames (Test)	QA Pairs (Test)	Format
	wea.	time.	con.	road.	area.	mark.	light.	sign.						
BDD-X	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	-	-	textual
BDD-OIA	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✓	-	-	textual
NuScenes-QA	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	36114	83337	textual
Talk2Car	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✓	1.8k	2447	textual
nuPrompt	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	36k	6k	textual
DRAMA	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✓	-	-	textual
Rank2Tell	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✓	-	-	partially structured
DriveLM	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	4794	15480	partially structured
DriveVLM	✓	✓	✗	✓	✗	✗	✗	✗	✓	✓	✓	-	-	partially structured
NuScenes-S	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	6019	18057	structured

TABLE I: Comparison among benchmark datasets for autonomous driving.

view of the driving scenario, addressing the often overlooked or insufficiently represented aspects in many existing datasets [4], [34], [28], [29], [33]. The scene description in NuScenes-S is structured and concise, which includes the following key elements: { *Weather, Traffic condition, Driving area, Traffic light, Traffic sign, Road condition, Lane markings, Time*}.

- **Weather**: Weather conditions play a crucial role in driving, as adverse weather can reduce visibility and alter road conditions, ultimately leading to more cautious driving decisions. Weather conditions include {*sunny, rainy, snowy, foggy, cloudy*}.
- **Traffic condition**: Different traffic conditions will introduce different driving challenges, traffic congestion will affect the driver's speed and decision-making. Traffic conditions include {*low, moderate, moderate*}.
- **Driving area**: Each driving area has its own characteristics, such as intersections and junctions that pose more challenges for turning and lane changing decisions. Driving areas include {*intersection, junction, roundabout, residential, crosswalk, parking lot*}.
- **Traffic light**: Traffic lights are important traffic control devices that regulate the flow of traffic. The state of traffic lights will affect the driver's decision-making. Traffic lights include {*green, yellow, red*}.
- **Traffic sign**: Traffic signs provide important information for drivers changing driving behavior to follow the rules. Traffic signs include {*speed limit, stop, yield, no entry, no parking, no stopping, no u-turn, no left turn, no right turn, no overtaking, one way*}.
- **Road condition**: Road conditions are critical for driving safety: construction zones require caution, and wet or icy roads require slower speeds and longer following distances. Road conditions include {*smooth, rough, wet, icy, construction*}.
- **Lane markings**: Lane markings provide directional guidance to guide drivers' driving decisions. Lane markings include {*right turn, left turn, straight, straight and right turn, straight and left turn, u-turn, left and u-turn, right and u-turn*}.
- **Time**: Time represents the time of day, driver tends to drive more cautiously at night due to reduced visibility. Time includes {*daytime, night*}.

B. Perception & Prediction

Identify some key objects and predict their future states are essential for a driver to make decisions. Most existing datasets [5], [34] describe these tasks in free-form language descriptions, which usually use a very long and verbose sentence or paragraph to describe a perception or prediction task while truly contains only a few key elements. To address this issue, we incorporated the perception and prediction tasks into the NuScenes-S dataset and managed them in a structured and concise manner to improve the efficiency and

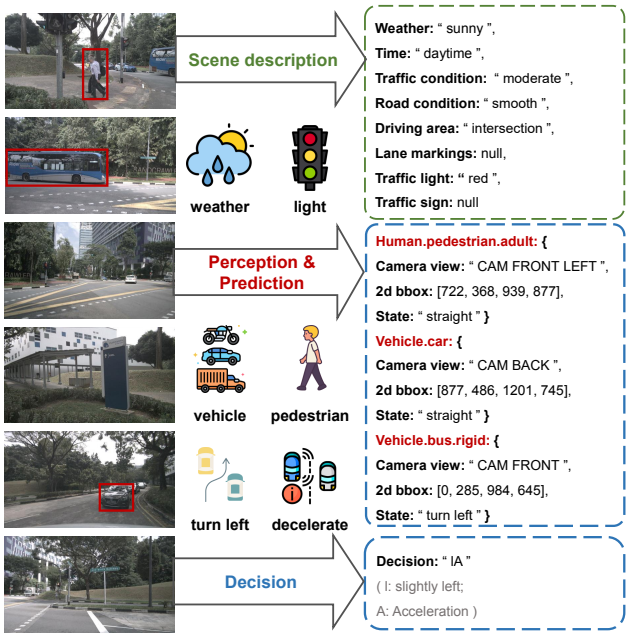


Fig. 3: An annotation example of the NuScenes-S dataset.

integration of VLMs. The perception and prediction tasks in NuScenes-S are structured as follows: {*Object:*{*Camera view, 2D bounding box, Future state*}}.

- **Object**: The object is the key element in the perception and prediction tasks, which includes the following attributes: {*Camera view, 2D bounding box, Future state*}.
- **Camera view**: The camera view of the object, which helps ego vehicle to identify the direction of the object in decision making. The camera view includes {*Front, Front left, Front right, Back, Back left, Back right*}.
- **2D bounding box**: The 2D bounding box of the object, which helps ego vehicle to locate the object in the camera view. The 2D bounding box consists of coordinates of the two diagonal vertices {*x1, y1, x2, y2*}.
- **Future state**: The future state of the object, the ego vehicle makes driving decisions based on the future state of the object. The future state includes {*Straight, Turn left, Turn right, Slightly left, Slightly right, Stop, Idle*}.

C. Decision

Make decisions based on the perception and prediction tasks is the final and critical step for a driver to drive safely. Current method rely on linguistic descriptions to describe the decision-making process that limit the integration of VLMs into autonomous systems. To address this issue, we treat the decision-making task as visual action reasoning thus convert the decision-making task into VLA task through defining some ruled-based actions similar to modular driving system. The decision-making task in NuScenes-S is structured as follows: {*Decision:* {*Lateral movement, Longitudinal movement*}}.

- **Decision**: The decision is a safe driving action that the ego vehicle could take based on the perception and prediction tasks, which includes the following attributes: {*Lateral movement, Longitudinal movement*}.
- **Lateral movement**: The lateral movement of the vehicle, which includes {*Turn left(L), Turn right(R), Slightly left(l), Slightly right(r), Straight(S)*}.
- **Longitudinal movement**: The longitudinal movement of the vehicle, which includes {*Accelerate(A), Decelerate(D), Cruising(C), Idle(I)*}.

D. Dataset Construction

In order to construct a high-quality structured benchmark dataset, we construct the datasets with a tiered and comparative optimization manner through holistic integration of rule-based annotation, VLM annotation, and human refinement, as is shown in Fig. 2. Specifically, in scene description, we first annotation scene information through GPT and human annotators, then we use compare the results of GPT and human annotators to find the difference and refine the annotations by human annotators. Similarly, in perception and prediction tasks, we first define some rules to extract key objects then we use VLMs and human annotators to annotation the key objects synchronously. Subsequently, through comparative optimization and human refinement to ensure the quality of the dataset. Finally, the related information of key objects could be extracted directly from NuScenes dataset. Finally, The decision task is annotated rule-based and human annotators to get initial annotations, then further refined by human annotators with comparative optimization. It is worth noting that by strategically organizing the annotation sequence, partial parallelization of the annotation tasks can be achieved, thereby improving annotation efficiency. On the other hand, by combining multiple annotation methods and employing contrastive optimization, the arbitrariness of relying on a single annotation method is avoided, further enhancing the quality of the dataset.

IV. FastDrive

The overview of the FastDrive is shown in Fig. 4. FastDrive is a compact VLM for end-to-end autonomous driving with parameters of 0.9B, significantly lower than current methods. The model follows the "ViT-Adapter-LLM" architecture widely used in various MLLM studies [11], [12] but introduced an optional TokenPacker module that reduce the number of visual tokens to improve the inference speed. Moreover, we fine tune the model by chaining autonomous driving tasks into a reasoning process, aiming to accelerate the model's learning of the relationships between these tasks and improve the model's performance on the NuScenes-S benchmark.

A. Vision Encoder

The backbone of the Vision Encoder is a Vision Transformer (ViT) based on Intern ViT-300M [35], which is distilled from the teacher model Intern ViT-6B [13]. The ViT backbone consists of a stack of 24 Transformer blocks with 16 heads, and the hidden size of the model is 1024 with 0.3B parameters. It can achieve a competitive performance on various vision-language tasks while maintaining a relatively small number of parameters by incrementally pre-training the model on large-scale datasets. As is shown in Fig. 4, the

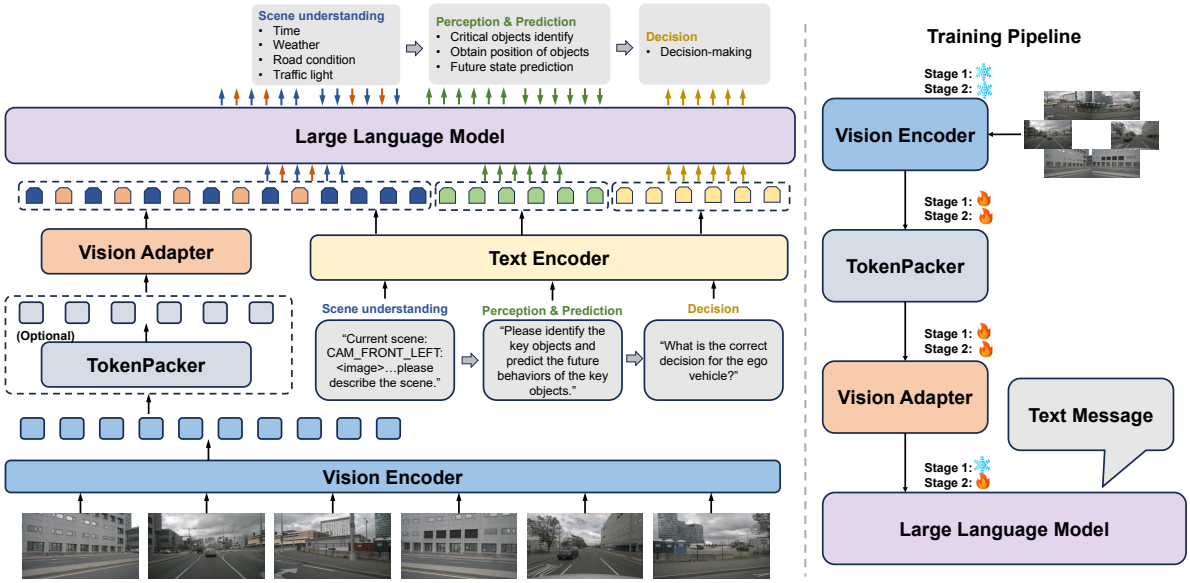


Fig. 4: The framework of the FastDrive model for end-to-end autonomous driving.

ViT backbone takes the input images of six different views, including front, front left, front right, back, back left, and back right, and extracts visual features from the images. The visual features are then projected into feature space of LLM by an MLP adapter. Additionally, we introduce an optional TokenPacker module to reduce the number of visual tokens, which can improve the inference speed of the model while maintaining the competitive performance.

B. LLM Agent

As illustrated in Fig. 4, the LLM plays a "brain" role in the FastDrive model throughout the driving process, which takes the visual features from the Vision Encoder and the structured language instructions as input and generates the scene description, identifys key objects, predicts their future states, and makes driving decisions in a chain of thought (CoT) manner. Specifically, we choose Qwen2.5 [36] as the LLM agent in the FastDrive model, which is a small LLM model with 0.5B parameters. Qwen2.5 has shown competitive performance on a wide range of benchmarks evaluating language understanding, reasoning, etc. It has achieved a significant improvement in instruction following, understands structured data and generates structured outputs. Thus, we select Qwen2.5 as the LLM agent in the FastDrive baseline model.

V. Experiments

A. Implementation Details

The experiments are conducted on 8 NVIDIA RTX 4090 GPUs. The FastDrive model is trained with a batch size of 1 for 10 epochs using the Adam optimizer with an initial learning rate of 1e-4, and the learning rate is decayed by a factor of 0.05. The experiments are conducted on the NuScenes-S dataset, which contains about 102K QA pairs in total. The dataset is split into 84K training QA pairs, 18K test QA pairs. The evaluation metrics include Language metrics, Average Precision (AP), Recall, Precision and Decision Accuracy for perception, prediction, and decision-making tasks.

B. Quantitative Results

Scene Understanding. The TABLE. II show the performance of scene understanding on the NuScenes-S dataset, the results demonstrate that the FastDrive model achieves competitive performance on the structured benchmark dataset. In TABLE. III, we compare the models' performance in perception, prediction, and decision-making tasks. The DriveLM model excels in perception with higher language metrics but lower accuracy, while FastDrive outperforms in prediction and decision-making with higher accuracy. This raises the issue that language evaluation metrics may not be suitable for assessing autonomous driving tasks, as they primarily measure fluency and coherence, which are important for natural language processing but do not capture the practical aspects of autonomous driving. These metrics focus on how well the generated text flows or aligns with human expectations, but they fail to evaluate the model's functional correctness in decision-making, perception, and real-world performance. In autonomous driving, what matters most is how effectively the model interprets sensory data and makes safe, accurate driving decisions, aspects that may be challenging to fully capture with verbose language descriptions.

Perception & Prediction & Decision. Additionally, it's worth noting that the language evaluation metrics get worse in perception tasks, which may raise another current VLMs may further improve the reasoning capabilities since the perception tasks are more challenging with complex and multimodal reasoning compared to the scene understanding. Moreover, the final task of end-to-end autonomous driving is generating safe and reasonable driving decisions, which is the most critical and challenging task for VLMs. From the TABLE. III, the FastDrive model achieves the best performance in decision-making tasks with the highest accuracy metrics. However, we also observed that the decision accuracy is relatively low. Further analysis revealed that the proportion of

Method	Language						Accuracy (%)							
	BLEU_1	BLEU_2	BLEU_3	BLEU_4	ROUGE_L	CIDEr	weather	time	traffic	road	area	mark	light	sign
DriveLM	82.70	76.51	70.41	65.05	83.93	5.30	85.47	99.91	76.30	83.85	74.96	81.49	85.57	83.90
FastDrive$_{64}$	80.49	77.66	72.77	68.06	60.53	3.58	93.35	99.81	78.08	86.57	75.98	**82.31**	**88.22**	85.85
FastDrive$_{256}$	**86.77**	**81.09**	**75.34**	**70.36**	**87.24**	**6.20**	**94.13**	**99.95**	**78.15**	**87.66**	**76.49**	82.06	87.74	**87.64**

TABLE II: Performance of scene description on the NuScenes-S dataset. **Bold** indicates the best performance. FastDrive$_{64}$ (with TokenPacker) and FastDrive$_{256}$ are the FastDrive models with 64 and 256 tokens, respectively. The same applies to the following tables.

Method	Perception								Prediction	Decision			
	BLEU_1	BLEU_2	BLEU_3	BLEU_4	ROUGE_L	CIDEr	AP	Recall	State	Dec	Dec(s)	Lat	Lon
DriveLM	**34.82**	**29.59**	**23.23**	**17.45**	**35.31**	0.74	0.21	0.30	0.36	0.28	0.59	0.72	0.35
FastDrive$_{64}$	26.07	15.17	8.86	4.25	34.37	**0.75**	0.31	0.45	**0.44**	0.38	**0.63**	0.74	0.45
FastDrive$_{256}$	26.48	15.23	9.11	4.75	34.77	0.61	**0.37**	**0.53**	**0.44**	**0.39**	**0.63**	**0.76**	**0.46**

TABLE III: Performance of perception, prediction, and decision-making tasks on the NuScenes-S dataset. DEC represents the accuracy of decision results that are consistent with the ground truth. Dec(s) represents the proportion of safe decisions, including those that match the ground truth as well as those that deviate from the ground truth but are still considered safe.

Method	Params	Trainable	Memory (GB)	FPS
DriveLM‡	3.955B	12.9M	14.43	0.20
DriveLM	3.955B	12.9M	14.43	0.36
FastDrive$_{64}$‡	0.9B	8.79M	1.97	2.86
FastDrive$_{256}$‡	0.9B	8.79M	1.97	2.11
FastDrive$_{64}$	0.9B	8.79M	1.97	4.85
FastDrive$_{256}$	0.9B	8.79M	1.97	4.01

TABLE IV: Comparison of model parameters, trainable parameters, FLOPs, and inference speed (FPS) for different models. ‡ indicates that the model is tested on the DriveLM dataset.

Scene	FastDrive				FastDrive w/o			
	Dec	Dec(s)	Lat	Lon	Dec	Dec(s)	Lat	Lon
weather	0.35	0.55	0.79	0.44	0.33	0.51	0.78	0.43
time	0.38	0.59	0.84	0.45	0.37	0.58	0.86	0.43
traffic	0.44	0.69	0.85	0.47	0.40	0.66	0.82	0.47
road	0.40	0.64	0.78	0.47	0.39	0.64	0.78	0.46
area	0.33	0.53	0.66	0.44	0.28	0.51	0.66	0.39
mark	0.39	0.64	0.77	0.47	0.39	0.63	0.77	0.46
light	0.39	0.69	0.57	0.48	0.42	0.63	0.68	0.51
sign	0.42	0.65	0.68	0.53	0.40	0.64	0.70	0.51

TABLE V: Ablation studies on the impact of scene annotations on driving decisions. w/o indicates the ablation study.

safe decisions is relatively high, indicating that the Vision-Language Model (VLM) tends to favor more conservative decisions. In addition, the accuracy of lateral (horizontal) decisions is higher than that of longitudinal (vertical) decisions, reflecting that longitudinal decision-making may be inherently more challenging.

Inference Acceleration We conduct comparative latency analysis across models in TABLE. IV. Experimental results demonstrate that FastDrive achieves 4.85 FPS inference speed while maintaining competitive performance on the NuScenes-S benchmark, representing a 13.5× acceleration over DriveLM's 0.36 FPS baseline. This efficiency stems from three synergistic optimizations: (1) Architectural compactness reduces computational overhead (0.9B vs. 3.96B parameters); (2) Systematic conversion of unstructured linguistic inputs into structured formats via NuScenes-S, eliminating redundant semantic processing; (3) Visual token compression reduces the number of visual tokens, further improving inference efficiency. While current implementation employs basic token pruning strategies, advanced visual compression architectures present promising directions for future investigation.

C. Ablation Studies

To evaluate the impact of scene annotation information on driving decisions, we design a comprehensive set of ablation experiments to observe how the absence of each factor influences the model's decision-making performance. Specifically, we perform a series of fine-tuning experiments, where we systematically remove individual types of scene annotation elements. Then we compare the performance of these ablated models with the fully annotated model in corresponding challenging scenarios, providing a detailed analysis of how different types of scene information contribute to the model's decision-making capabilities. The results are shown in TABLE. V and Fig. 5. The results show that the FastDrive model with complete scene annotations achieves better performance in driving decisions than the FastDrive model without specific scene annotations in challenging scenarios, which indicates that the scene annotations are beneficial for the model to make safe and reasonable driving decisions.

It's worth noting that in the traffic light ablation experiment, the ablated model slightly outperformed the complete model. This can be attributed to the logical complexity introduced by traffic lights and the conservative nature of the model. From the results, we observe that the model more tends to adopt overly conservative decisions when traffic light information is provided. As is illustrated in Fig. 5, the ego vehicle tends to turn left when the traffic light is red, which is a safe and reasonable drive decision. The model tends to adopt overly conservative decisions to ensure safety when capture the traffic light information while the ablation model relies more directly on dynamic scene context and the behavior of surrounding traffic participants, allowing it to make decisions that align more closely with the actual ground truth. In all, the results show that traffic lights do

Fig. 5: Examples of ablation studies on the impact of scene annotations on driving decisions. The red decision represents a decision that is not consistent with the ground truth.

indeed impact the model's decision-making, highlighting a potential research direction for the efficient integration of scene information to strike a balance between safety and accuracy in autonomous driving.

VI. Conclusion

In this work, we introduce the NuScenes-S dataset, a structured benchmark dataset for autonomous driving, which follows the human-like reasoning process across perception, prediction, and decision-making tasks. The NuScenes-S address the limitations of redundancy and synonymous expressions caused by free-form and lengthy language descriptions in existing datasets through structured labeling. This approach reduces the complexity of handling unstructured information, allowing the model to process and interpret data more effectively, leading to more efficient decision-making. We also present the FastDrive, a compact VLM for end-to-end autonomous driving, which achieves competitive performance on the NuScenes-S dataset with faster inference speed and fewer parameters on NuScenes-S dataset. This highlights the potential of small-parameter models in structured benchmark datasets. Moreover, we conduct extensive experiments analysis the impact of scene annotations on driving decisions, which demonstrates that the scene annotations are beneficial for the model to make safe and reasonable driving decisions. We believe that the NuScenes-S dataset and the FastDrive model will serve as a valuable resource for future research in autonomous driving and structured benchmark datasets.

REFERENCES

[1] R. Esteban, L. Jannik, N. Uhlemann, and M. Lienkamp, "Scenario understanding of traffic scenes through large visual language models," 2025. [Online]. Available: https://arxiv.org/abs/2501.17131

[2] X. Zhou, M. Liu, E. Yurtsever, B. L. Zagar, W. Zimmer, H. Cao, and A. C. Knoll, "Vision language models in autonomous driving: A survey and outlook," 2024. [Online]. Available: https://arxiv.org/abs/2310.14414

[3] S. Xie, L. Kong, Y. Dong, C. Sima, W. Zhang, Q. A. Chen, Z. Liu, and L. Pan, "Are vlms ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives," 2025. [Online]. Available: https://arxiv.org/abs/2501.04003

[4] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, J. Beißwenger, P. Luo, A. Geiger, and H. Li, "Drivelm: Driving with graph visual question answering," 2025. [Online]. Available: https://arxiv.org/abs/2312.14150

[5] 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," 2024. [Online]. Available: https://arxiv.org/abs/2402.12289

[6] J. Mei, Y. Ma, X. Yang, L. Wen, X. Cai, X. Li, D. Fu, B. Zhang, P. Cai, M. Dou, B. Shi, L. He, Y. Liu, and Y. Qiao, "Continuously learning, adapting, and improving: A dual-process approach to autonomous driving," 2024. [Online]. Available: https://arxiv.org/abs/2405.15324

[7] T. Qian, J. Chen, L. Zhuo, Y. Jiao, and Y.-G. Jiang, "Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario," 2024. [Online]. Available: https://arxiv.org/abs/2305.14836

[8] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for self-driving vehicles," 2018. [Online]. Available: https://arxiv.org/abs/1807.11546

[9] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," 2020. [Online]. Available: https://arxiv.org/abs/1903.11027

[10] Z. Chen, W. Wang, Y. Cao, Y. Liu, and Z. Gao, "Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling," 2025. [Online]. Available: https://arxiv.org/abs/2412.05271

[11] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023. [Online]. Available: https://arxiv.org/abs/2304.08485

[12] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," 2024. [Online]. Available: https://arxiv.org/abs/2310.03744

[13] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, B. Li, P. Luo, T. Lu, Y. Qiao, and J. Dai, "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," 2024. [Online]. Available: https://arxiv.org/abs/2312.14238

[14] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," 2023. [Online]. Available: https://arxiv.org/abs/2304.10592

[15] X. Ding, J. Han, H. Xu, X. Liang, W. Zhang, and X. Li, "Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models," 2024. [Online]. Available: https://arxiv.org/abs/2401.00988

[16] B. Lin, Z. Tang, Y. Ye, J. Huang, J. Zhang, Y. Pang, P. Jin, M. Ning, J. Luo, and L. Yuan, "Moe-llava: Mixture of experts for large vision-language models," 2024. [Online]. Available: https://arxiv.org/abs/2401.15947

[17] Y. Zhang, J. Lu, and N. Jaitly, "The entity-deduction arena: A playground for probing the conversational reasoning and planning capabilities of LLMs," 2024. [Online]. Available: https://openreview.net/forum?id=PfrpYGKGPL

[18] Z. Yang, X. Jia, H. Li, and J. Yan, "Llm4drive: A survey of large language models for autonomous driving," 2024. [Online]. Available: https://arxiv.org/abs/2311.01043

[19] L. Wen, D. Fu, X. Li, X. Cai, T. Ma, P. Cai, M. Dou, B. Shi, L. He, and Y. Qiao, "Dilu: A knowledge-driven approach to autonomous driving with large language models," 2024. [Online]. Available: https://arxiv.org/abs/2309.16292

[20] D. Fu, X. Li, L. Wen, M. Dou, P. Cai, B. Shi, and Y. Qiao, "Drive like a human: Rethinking autonomous driving with large language models," 2023. [Online]. Available: https://arxiv.org/abs/2307.07162

[21] S. Xie, L. Kong, Y. Dong, C. Sima, W. Zhang, Q. A. Chen, Z. Liu, and L. Pan, "Are vlms ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives," 2025. [Online]. Available: https://arxiv.org/abs/2501.04003

[22] H. Shao, Y. Hu, L. Wang, S. L. Waslander, Y. Liu, and H. Li, "Lmdrive: Closed-loop end-to-end driving with large language models," 2023. [Online]. Available: https://arxiv.org/abs/2312.07488

[23] W. Wang, J. Xie, C. Hu, H. Zou, J. Fan, W. Tong, Y. Wen, S. Wu, H. Deng, Z. Li, H. Tian, L. Lu, X. Zhu, X. Wang, Y. Qiao, and J. Dai, "Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving," 2023. [Online]. Available: https://arxiv.org/abs/2312.09245

[24] 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," 2024. [Online]. Available: https://arxiv.org/abs/2410.22313

[25] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013. [Online]. Available: https://doi.org/10.1177/0278364913491297

[26] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, and V. Patnaik, "Scalability in perception for autonomous driving: Waymo open dataset," 2020. [Online]. Available: https://arxiv.org/abs/1912.04838

[27] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," 2020. [Online]. Available: https://arxiv.org/abs/1903.11027

[28] T. Deruyttere, S. Vandenhende, D. Grujicic, L. Van Gool, and M.-F. Moens, "Talk2car: Taking control of your self-driving car," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019. [Online]. Available: http://dx.doi.org/10.18653/v1/D19-1215

[29] D. Wu, W. Han, T. Wang, Y. Liu, X. Zhang, and J. Shen, "Language prompt for autonomous driving," 2023. [Online]. Available: https://arxiv.org/abs/2309.04379

[30] T. Qian, J. Chen, L. Zhuo, Y. Jiao, and Y.-G. Jiang, "Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario," 2024. [Online]. Available: https://arxiv.org/abs/2305.14836

[31] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for self-driving vehicles," 2018. [Online]. Available: https://arxiv.org/abs/1807.11546

[32] Y. Xu, X. Yang, L. Gong, H.-C. Lin, T.-Y. Wu, Y. Li, and N. Vasconcelos, "Explainable object-induced action decision for autonomous vehicles," 2020. [Online]. Available: https://arxiv.org/abs/2003.09405

[33] S. Malla, C. Choi, I. Dwivedi, J. H. Choi, and J. Li, "Drama: Joint risk localization and captioning in driving," 2022. [Online]. Available: https://arxiv.org/abs/2209.10767

[34] E. Sachdeva, N. Agarwal, S. Chundi, S. Roelofs, J. Li, M. Kochenderfer, C. Choi, and B. Dariush, "Rank2tell: A multimodal driving dataset for joint importance ranking and reasoning," 2023. [Online]. Available: https://arxiv.org/abs/2309.06597

[35] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu, L. Gu, X. Wang, Q. Li, Y. Ren, Z. Chen, J. Luo, J. Wang, T. Jiang, B. Wang, C. He, B. Shi, X. Zhang, H. Lv, Y. Wang, W. Shao, P. Chu, Z. Tu, T. He, Z. Wu, H. Deng, J. Ge, K. Chen, K. Zhang, L. Wang, M. Dou, L. Lu, X. Zhu, T. Lu, D. Lin, Y. Qiao, J. Dai, and W. Wang, "Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling," 2025. [Online]. Available: https://arxiv.org/abs/2412.05271

[36] Qwen, :, A. Yang, B. Yang, B. Zhang, B. Hui, and B. Zheng, "Qwen2.5 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2412.15115