Reason2Drive: Towards Interpretable and Chain-based Reasoning for Autonomous Driving

Ming Nie[1], Renyuan Peng[1], Chunwei Wang[2], Xinyue Cai[2], Jianhua Han[2], Hang Xu[2⋆], and Li Zhang[1⋆]

[1] School of Data Science, Fudan University
[2] Huawei Noah's Ark Lab

https://github.com/fudan-zvg/reason2drive

Abstract. Large vision-language models (VLMs) have garnered increasing interest in autonomous driving areas, due to their advanced capabilities in complex reasoning tasks essential for highly autonomous vehicle behavior. Despite their potential, research in autonomous systems is hindered by the lack of datasets with annotated reasoning chains that explain the decision-making processes in driving. To bridge this gap, we present Reason2Drive, a benchmark dataset with over 600K video-text pairs, aimed at facilitating the study of interpretable reasoning in complex driving environments. We distinctly characterize the autonomous driving process as a sequential combination of *perception*, *prediction*, and *reasoning* steps, and the question-answer pairs are automatically collected from a diverse range of open-source outdoor driving datasets, including nuScenes, Waymo and ONCE. Moreover, we introduce a novel aggregated evaluation metric to assess chain-based reasoning performance in autonomous systems, addressing the reasoning ambiguities of existing metrics such as BLEU and CIDEr. Based on the proposed benchmark, we conduct experiments to assess various existing VLMs, revealing insights into their reasoning capabilities. Additionally, we develop an efficient approach to empower VLMs to leverage object-level perceptual elements in both feature extraction and prediction, further enhancing their reasoning accuracy. Extendable experiments demonstrate the supportive effect of Reason2Drive towards visual reasoning and downstream planning tasks.

1 Introduction

Modern autonomous driving systems face challenges related to generalization issues across diverse scenarios, which is often attributed to the reliance on empirical and intricate rules involved in decision-making. To reduce dependence on such rules, recent end-to-end approaches [5, 20] have been developed to derive control signals directly from sensor inputs, treating the system as a black

⋆ Li Zhang (lizhangfd@fudan.edu.cn) and Hang Xu (chromexbjxh@gmail.com) are the corresponding authors.

box that requires extensive data for training. However, this approach tends to obscure the underlying logic of decisions, complicating failure diagnosis in real-world applications. In contrast, Large Vision-Language Models (VLMs) offer a promising alternative, potentially enhancing interpretability and generalization for these systems. With their broad world knowledge and advanced reasoning abilities, as illustrated in Fig. 1(a), VLMs have the potential to provide a more thorough understanding and explicit explanation for reliable decision-making. Nonetheless, existing works [35, 40] primarily focused on the straightforward adaptation of question-answering tasks to the autonomous driving; how to exploit VLMs to facilitate the reasoning abilities of autonomous systems is still under exploration.

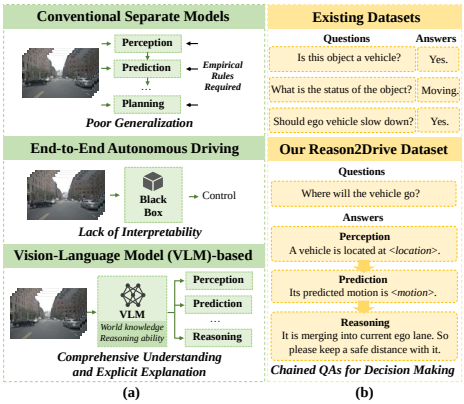

Fig. 1: (a) Different decision-making processes in autonomous driving. (b) Language-based dataset comparison.

One reason that hinders the research in this field lies in the scarcity of datasets, especially those chained-based reasoning labels that elucidate the decision-making process. Most existing datasets [11, 35, 41] often oversimplify the complex processes of driving into straightforward question-answering tasks with only a few specific tasks covered. As depicted in Fig. 1(b), they typically provide closed-form annotations constrained to boolean (i.e., yes or no) answers or limited multiple-choice responses (e.g., stopped, parked, and moving). However, autonomous driving transcends a simplistic QA process. It encompasses a multi-step approach involving *perception*, *prediction*, and *reasoning*, each of which plays an indispensable role in the decision-making. Therefore, it is crucial to introduce a novel benchmark annotated with detailed decision-making reasoning for assessing the reasoning abilities of current VLMs.

To this end, we introduce Reason2Drive, a new benchmark comprising over 600K video-text pairs, characterized by intricate driving instructions and a series of perception, prediction and reasoning steps. Our benchmark builds upon widely-used open-source driving datasets including nuScenes [2], Waymo [37], and ONCE [28], utilizing an extensible annotation schema. Specifically, we structure the comprehensive annotations into object-centric database and integrate it into manually crafted templates to create paired data for VLMs at both object and scenario levels. To enhance diversity, GPT-4 and manual instructions are employed for verification and enrichment purposes. Notably, Reason2Drive is the most extensive dataset available to date, outperforming existing datasets in scale and the complexity of reasoning chains included, which is a distinctive attribute not present in other datasets.

Furthermore, we observe a fundamental flaw in the current evaluation of VLMs on autonomous driving tasks, due to the inherent reasoning ambiguities of traditional caption-based metrics like BLEU [31] and CIDEr [39]. These metrics mainly measure text generation from a holistic perspective, without considering the causal relationship between the reasoning steps and the final conclusion. For example, when the VLM suggests ego vehicle turning left, we cannot ascertain from these metrics whether its reasoning steps effectively support the final decision. To address this issue, we propose a new aggregated evaluation metric, ADRScore, specifically designed to measure chain-based reasoning performance in autonomous systems, which aims to resolve the reasoning ambiguities associated with current metrics.

Utilizing the proposed benchmark, we undertake experiments to assess various existing VLMs, thereby unveiling valuable insights into their reasoning capabilities. We find that most methods struggle to effectively leverage perceptual priors, resulting in subpar reasoning performance. Additionally, constrained by the language model functioning solely as a decoder, these methods often fail to deliver accurate perceptual results, which is a crucial component for verifying a model's spatial reasoning capability. To alleviate this dilemma, we present a simple yet efficient framework, augmenting existing VLMs with two new components: a prior tokenizer and an instructed vision decoder, which aim to bolster the models' visual localization capabilities within the encoder and decoder, respectively. Extendable experiments demonstrate the supportive effect of Reason2Drive towards visual reasoning and downstream planning tasks.

The contributions of this paper are summarized as follows: **(i)** We publish a novel visual instruction tuning dataset aimed at facilitating interpretable and chain-based reasoning autonomous systems. **(ii)** We introduce a novel evaluation metric, ADRScore, to assess chain-based reasoning performance in autonomous driving, effectively addressing the reasoning ambiguities present in existing metrics. **(iii)** We conduct experiments to assess a range of existing VLMs, revealing valuable insights into their reasoning capabilities. **(iv)** To address the challenges posed by inefficient prior feature extraction and inaccurate perceptual predictions, we introduce an efficient approach to integrate these elements into VLMs. This results in a substantial improvement in reasoning accuracy and provides remarkable support for downstream planning tasks. Our method surpasses all baselines, notably achieving impressive generalization in unseen scenarios.

2 Related Work

Multimodal large language model. The current state of large language models provides remarkable abilities in natural language understanding and generation ([6, 7, 30, 38]). Inspired by the potential of large language models, a multitude of multimodal models has emerged, intended to enhance these models' capabilities in achieving multi-modal comprehension. Blip-2 [24] aligns visual and language features by utilizing a learnable Q-former. LLaVA [25] and MiniGPT-4 [46] initially align image-text features and then proceed with in-

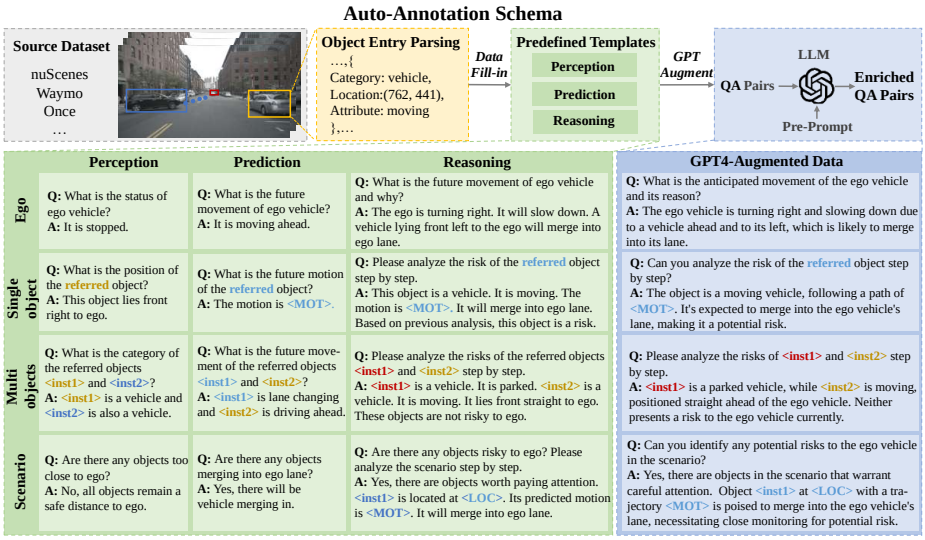

Fig. 2: Schema of Our Reason2Drive Dataset. The upper part illustrates the pipeline for the automated construction of datasets. The lower part shows detailed instances of perception, prediction, and reasoning, accompanied by outcomes after applying GPT-4 for data augmentation. The special tokens hold distinct definitions: `<Inst*>` represents a specified instance, `<MOT>` signifies a forecasted sequence of trajectory coordinates, and `<LOC>` denotes positional coordinates. The colors associated with these tokens correspond to the highlighted objects in the upper-left image's boxes.

struction tuning. Additionally, Video-LLaMA [44] and ImageBind-LLM [18] integrate multiple modalities into the input, aligning features from various sources like images, videos, audio, and point clouds, consolidating them into the space of language features. Kosmos-2 [33] and Shikra [4] perform object detection based on instructions and also accomplish grounded visual question answering. DetGPT [34] connects a fixed multi-modal LLM with a customizable detector based on user instructions. LISA [23] efficiently embeds segmentation abilities into multi-modal LLMs, showcasing self-reasoning for current perception systems. The previous works have demonstrated that current large-scale multimodal models can achieve cross-modal alignment, enabling comprehension and inference towards images and more. These models can not only perform perceptual tasks like detection but also accomplish preliminary reasoning tasks.

Vision language tasks in autonomous driving. Currently, VLMs have demonstrated robust capabilities in scene perception and understanding. Extensive efforts have been dedicated to the realm of autonomous driving, leveraging VLM to achieve comprehensive scene understanding and perform diverse tasks [13, 16, 29, 42]. Simultaneously, substantial works are in progress to create datasets and models tailored to various tasks. Talk2Car [11] proposes the first object referral dataset for grounding commands for self-driving cars in free natural language into the visual context. But it exclusively contains information about visible objects. While DRAMA [27] outlines the overall scene risk, it lacks pre-

cise perception annotation. NuPrompt [41] and Refer-KITTI [40] offer language prompt sets for driving scenes but primarily concentrate on multi-object tracking tasks. NuScenesQA [35] and DriveLM [9] build visual question-answering (VQA) datasets for scenario understanding. However, their primary emphasis is on the perceptual information in the scene, lacking annotations for the analysis and complex reasoning of the entire scenario. To address the limitations of existing works, we construct a thorough dataset covering perception, prediction, and complex reasoning, additionally with an improved vision-language model for better analyzing autonomous driving scenarios.

3 Reason2Drive Dataset

We introduce Reason2Drive, a dataset that comprises comprehensive driving instructions and a chain-based reasoning framework for decision-making. Our dataset is characterized by the following key aspects:

- **Quantity**: It stands out as the largest language-based driving dataset available, collated from prominent publicly accessible datasets worldwide.
- **Quality**: Reason2Drive offers a more precise representation of driving activities, including *perception*, *prediction* and *reasoning*, with a reliable auto-annotation schema for data collection.
- **Diversity**: (i) The dataset exhibits a broader range of scenes, encompassing both object-level and scenario-level data. This diversity includes object types, visual and motion attributes, object locations, and relationships relative to the ego-vehicle. (ii) It includes more intricate question-answer pairs, enhanced by GPT-4, along with longer text passages featuring step-by-step reasoning.
- **Protocols**: A novel evaluation metric is introduced to assess the reasoning capabilities of VLMs. Different from those widely used in the NLP community, it takes into account not only perception results but also reasoning ambiguities, providing a more holistic evaluation of the VLM's reasoning capacity for autonomous driving scenarios.

Further details regarding the data collection process, statistical data analysis, and benchmark protocols are provided in the subsequent section.

3.1 Dataset collection

As illustrated in Fig. 2, we employ an extensible annotation schema, constructing data in the forms of question-answer pairs. Specifically, we first leverage a diverse array of publicly available datasets collected in different regions worldwide, including nuScenes, Waymo, and ONCE, and then parse their comprehensive annotations into object-centric database. The database is organized in key frames, and each frame stores object entries containing various details pertaining to its

Dataset	Description	Source Datasets
Talk2Car [11] ■	Object referral	nuScenes
CityFlow-NL [15] ■	Tracking & retrieval	CityFlow
CARLA-NAV [21] ■	Segmentation & prediction	CARLA Simulator
NuPrompt [41] ■	Multi-object tracking	nuScenes
NuScenes-QA [35] ■	Perception	nuScenes
Refer-KITTI [40] ■	Multi-object tracking	KITTI
Talk2BEV [12] ■	Visual understanding	nuScenes
DRAMA [27] ■	Risk localization	self-collected
Rank2Tell [36] ■	Risk localization & ranking	self-collected
Reason2Drive■	Perception Prediction Reasoning	nuScenes Waymo ONCE

Table 1: The comparison between our Reason2Drive dataset and other prompt-based datasets. ■ means dataset not published.

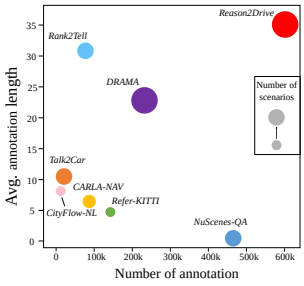

Fig. 3: Data quality comparison. Reason2Drive is larger in scale, richer in data content, and more diverse in scenarios.

driving actions, including location, category, attributes and more. Afterwards, the structured database is required to answer the manually crafted question templates, which are divided into different tasks (i.e., perception, prediction and reasoning) at both object-level and scenario-level. Subsequently, GPT-4 is involved for verification and enrichment purposes. The example of GPT augmented data and manual instructions are provided in the appendix.

Due to the complexity of autonomous driving activities, we categorize the tasks into three distinct groups to acquire diversified data: perception, prediction and reasoning. The specifics and distinctions of these three types of tasks are elaborated as follows:

- **Perception task** is designed to identify objects within the driving scenario, assessing the fundamental perceptual capabilities of VLMs in outdoor environments.
- **Prediction task** entails the prediction of future states of key objects within the perceptual range, challenging VLMs to infer the intentions of objects with video input.
- **Reasoning task** prompts the analysis of the current perceptual and predicted states step by step, requiring the deduction of reasoned inferences and decisions through a chain of thoughts (COT) approach.

For each task, we further categorize the data into object-level and scenario-level. In more detail,

- **Object-level** data is formatted to benchmark the foundational capabilities of specific objects. As for perception, we address the location and attributes of objects such as moving status and distance to ego, while for prediction, future motion and merging-in/out status are considered.
- **Scenario-level** data is organized from a global perspective towards driving environment and ego-driving instructions. It focuses on whether there is an object worth noting currently (perception), whether there is an object worth noting in the future (prediction) and why (reasoning). For example,

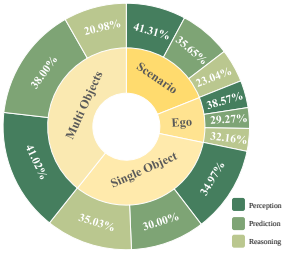

Fig. 4: Statistical distribution of different tasks in Reason2Drive.

Target \ Task	Perception (PE)	Prediction (PR)	Reasoning (RE)	Total
Ego vehicle	22629	17173	18868	58670
Single object	71882	61667	72006	205555
Multi objects	102012	94502	52175	248689
Scenario	49589	42795	27657	120041
Total	246112	216137	170706	632955

Table 2: The statistics of different tasks in Reason2Drive dataset.

as illustrated in Fig. 2, models are asked to identify distances, merging states and other risks from the whole scene. It verifies the agent's ability to perceive the entire driving scene rather than specifying objects, thus more challenging.

3.2 Dataset analysis

Tab. 1 and Fig. 3 demonstrate the comparison between our Reason2Drive dataset and existing benchmarks. It is noteworthy that our benchmark stands as the largest dataset to date, surpassing others in terms of both dataset size and the inclusion of extensive long-text chain-based reasoning references.

To further investigate the property of Reason2Drive dataset, we count the distribution and sample numbers of our dataset in Fig. 4 and Tab. 2. We split the dataset according to the task and target. The benchmark exhibits a balanced distribution, with multi-object tasks constituting the majority. Single-object and scenario-level questions are of similar quantities. The fewest questions are related to the ego-vehicle. Additionally, perception, prediction and reasoning questions are distributed as 39%, 34%, and 27%, respectively. More dataset details are provided in the appendix.

3.3 Benchmark protocol

It is worth noting that previous works [12, 27, 35] simply utilize metric scores widely used in the NLP community, including BLEU [31], CIDEr [39] and METEOR [1]. However, these metrics mainly measure text generation from a holistic perspective, without considering the causal relationship between the reasoning steps and the final conclusion. While they perform well in translation and captioning, their efficacy is limited when it comes to reasoning. Moreover, the existing evaluation system is confined to assessing the semantic quality of the results. It fails to effectively evaluate the quality of the perceived results, a crucial aspect supporting automatic driving decision-making. To address these dilemmas, inspired by [43] and [17], we develop a novel evaluation protocol, ADRScore, to measure the correctness of the reasoning chains towards autonomous driving.

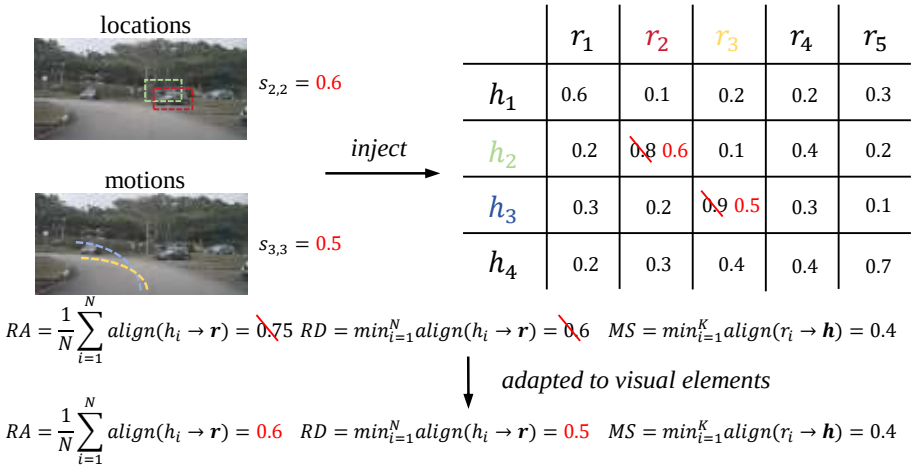

Fig. 5: Illustration of ADRScore and its visual adaptation. When substituting semantic similarities with actual geometric similarities, the score decreases.

Preliminary. To begin with, we denote the generated reasoning steps as hypothesis $\boldsymbol{h} = \{h_1, ..., h_N\}$, and the gold annotation as reference $\boldsymbol{r} = \{r_1, ..., r_K\}$.

At the core of reasoning metrics is the reasoning alignment vector from the N-step hypothesis h to the K-step reference:

$$align(\boldsymbol{h} \to \boldsymbol{r}) = \{\alpha_1, ..., \alpha_N\}, \tag{1}$$

where alignment value α_i represents the semantic similarity between the corresponding hypothesis step and the most similar reference step:

$$\begin{aligned} \alpha_i &= max_{j=1}^{K} s_{i,j}, \\ s_{i,j} &= cos(h_i, r_j). \end{aligned} \tag{2}$$

$\alpha_i \in [0, 1]$ explicitly measures the grounding of the step-wise reasoning with respect to the reference, and $cos(\cdot)$ denotes the cosine similarity between the corresponding sentence embeddings, which are extracted a pre-trained bert model. Based on the above reasoning alignment vector, we propose the following metrics to thoroughly measure the quality of reasoning steps.

Reasoning alignment. The most straightforward way to evaluate the correctness of the hypothesis reasoning chain is to compare the degree of overlap between the hypothesis and the reference. One way of doing that is to measure the reasoning alignment between them:

$$RA = \frac{1}{N}\sum_{i=1}^{N} align(h_i \to \boldsymbol{r}). \tag{3}$$

Redundancy. To find chains that contain information that is not required to solve the problem (i.e., redundant steps), we identify those hypothesis steps that

are least aligned with the reference steps. This metric punishes the chain with steps that are not required for the correct solution.

$$RD = min_{i=1}^{N} align(h_i \rightarrow \boldsymbol{r}). \tag{4}$$

Missing step. To identify steps that are missing from the hypothesis but could be required to solve the problem, we look at the alignment between reference and the hypothesis, similar to *Redundancy*. However, here we go through each step in the reference, and check if there is a similar step in the hypothesis:

$$MS = min_{i=1}^{K} align(r_i \rightarrow \boldsymbol{h}). \tag{5}$$

Finally, the aggregated metric score is the average of the above performance, which is:

$$ADRScore = \frac{1}{3}(RA + RD + MS). \tag{6}$$

Adapted to visual elements. To further adapt to the realistic driving process, we promote the above metric to the situation with visual elements, named ADRScore-S. Specifically, as illustrated in Fig. 5, when the hypothesis step h_i and reference step r_j contains visual elements, *i.e.*, the locations and motions predicted for further reasoning, the similarity score becomes:

$$s_{i,j} = \frac{\tau - M(h_i, r_j)}{\beta}, \tag{7}$$

where $M(\cdot)$ measures the mean square error between two perceptual elements. And τ and β are used to normalize it to $[0, 1]$ to match the distribution of semantic-level similarity. ADRScore-S more harshly measures the performance on spatial reasoning as the metric calculates the error of visual elements in spatial instead of text semantic. The latter is too lenient for visual predictions in language.

4 Methodology

In this section we introduce our framework in Sec. 4.1, followed by the training details provided in Sec. 4.2.

4.1 Model architecture

We observe that most VLMs struggle to effectively handle object-level perceptual information, including the input of visual priors and predictions of object locations, which are indispensable in autonomous driving scenarios. The limitation is primarily due to (i) the lack of a targeted tokenizer and (ii) decoder solely composed of a language model, resulting in subpar reasoning performance.

To address this challenge, as illustrated in Fig. 6, we introduce a straightforward yet effective framework that enhances existing VLMs with two new components: a prior tokenizer and an instructed vision decoder. Notably, the design

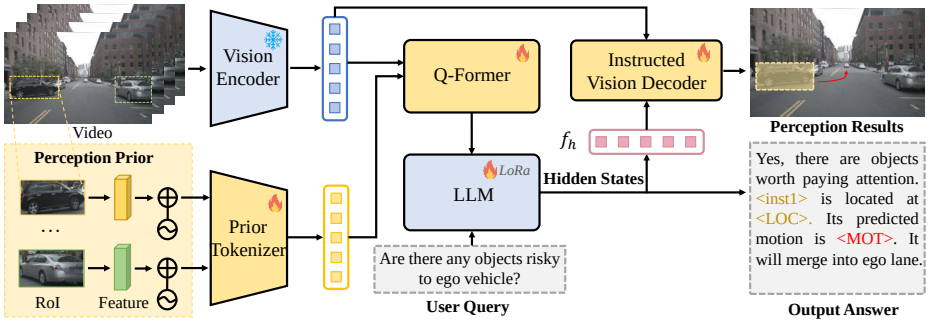

Fig. 6: The pipeline of our framework. The input video and perceptual priors are tokenized using the vision encoder and prior tokenizer. A Q-former then aligns them to the text's feature space. The LLM and instructed vision decoder predict answers with precise perception results for user queries. The highlighted yellow box and red curve in the perception results respectively represent the visualization of `<LOC>` and `<MOT>`.

of the components is not for detection but to aid interpretable visual reasoning based on available perception inputs. These modules aim to strengthen the capabilities of the model to utilize object-level perceptual elements in both extracting visual priors and generating perceptual predictions for visual reasoning.

Vision encoder. Our model accepts both video frames and text inputs, along with perceptual priors, and tokenizes them into embeddings. For a sequence of video frames $(V_1, V_2, ..., V_N)$, features are extracted using a pretrained Blip-2 visual encoder [24] F_v and aggregated through concatenation:

$$f_v = F_v(V_1) \oplus F_v(V_2) \oplus ... \oplus F_v(V_N). \tag{8}$$

Prior tokenizer. We propose a novel tokenization strategy tailored to taking advantage of visual cues. The motivation is grounded in the acknowledgement that extracting and aligning visual features is considerably simpler and more suitable compared to compelling the LLM to comprehend ambiguous positional descriptions. Direct textual input to the LLM may result in challenges such as information loss, as textual representation may not fully capture image details and context, especially in complex scenarios with dynamic object positions and velocities. To tackle this issue, we design a novel tokenizer F_p, implemented as a two-layer MLP, to independently extract local image features and positional embeddings from visual priors:

$$f_p = F_p(f_r + E(P)), \tag{9}$$

where f_r represents the region-level features extracted from the image-level features f_v according to the precise locations of perception priors P. These features are aligned to 7×7 size using the RoIAlign [14] operation and fused into a single embedding f_r. And $E(\cdot)$ is a positional encoding function mapping the geometry locations and motions into the same dimension of f_r.

LLM. After we tokenize the video and perception priors into embedding f_v and f_p, a projector Q (Q-former [24] in this work) is adopted to align the non-text features into textual domain:

$$f_q = Q(f_v, f_p). \tag{10}$$

Then, to generate the final text output, we utilize the LLM for further language processing with the extracted text embedding f_t:

$$\hat{y}_t = F(f_t, f_q). \tag{11}$$

Instructed vision decoder. Current works [11, 16] treat the LLM as a versatile tool to generate answers and inferences without intermediate reasoning steps, let alone considering the perceptions of the agent toward driving scenes. However, the perception ability of the agent towards driving scenarios is an indispensable part of a reliable driving procedure. Moreover, recent works [23] have demonstrated that, rather than training with textualized perceptual sequences, incorporating the perception abilities into the multi-modal LLM brings a significant improvement. To this end, inspired by [23], we integrate new perception capabilities into the multi-modal LLM. Specifically, we expand the original LLM vocabulary by introducing new tokens as placeholders, denoted as `<LOC>` and `<MOT>`, to signify the request for the perception output. When the LLM aims to generate a specific perception, the output $\hat{y}_t$ should include a designed token. We then extract the last-layer textual features corresponding to the specific token and apply an MLP projection layer to obtain the hidden embedding f_h. Finally, the textual embedding and visual features are fed into the instructed vision decoder to decode the predictions:

$$\hat{P} = D(f_v, f_h). \tag{12}$$

This module is comprised of a transformer decoder for features alignment [3] and task-specific heads designed to generate object locations and motions.

4.2 Training details

Training objectives. The model is trained end-to-end using the text generation loss $\mathcal{L}_{txt}$ and the perception output loss $\mathcal{L}_{per}$:

$$\mathcal{L} = \mathcal{L}_{txt} + \lambda_{per}\mathcal{L}_{per}, \tag{13}$$

where λ_{per} is the balancing term. Specifically, $\mathcal{L}_{txt}$ is the auto-regressive cross-entropy loss for text generation, and $\mathcal{L}_{per}$ encourages the instructed vision decoder to generate accurate locations and motions, which is similar to traditional detection loss and is employed with the combination of binary cross-entropy loss and MSE loss. More details are included in the appendix.

Tuning strategy. Our tuning strategy consists of two stages: the pre-training stage and the fine-tuning stage. In the pre-training stage, we initialize the weights from instructBLIP [10], including the pre-trained vision encoder, Q-former and

Table 3: Results of different models on the Reason2Drive validation set. We evaluate with ADRScore as well as captioning-based metrics.

Methods	LLM	Reasoning metric		Captioning metric			
		ADRScore	ADRScore-S	B@4	METEOR	ROUGE	CIDEr
Blip-2 [24]	OPT-2.7B [45]	0.296	0.162	0.361	0.249	0.443	0.174
	FlanT5-XL [8]	0.310	0.171	0.368	0.256	0.451	0.187
InstructBLIP [10]	FlanT5-XL	0.329	0.187	0.376	0.269	0.462	0.196
	Vicuna-7B [32]	0.351	0.214	0.408	0.294	0.484	0.211
MiniGPT-4 [46]	Vicuna-7B	0.338	0.203	0.396	0.286	0.475	0.219
Ours	FlanT5-XL	0.457	0.420	0.451	0.349	0.520	0.292
	Vicuna-7B	**0.463**	**0.432**	**0.457**	**0.356**	**0.529**	**0.298**

Table 4: Ablations on different combinations of training tasks.

Tasks			Reasoning metric		Captioning metric			
Perception	Prediction	Reasoning	ADRScore	ADRScore-S	B@4	METEOR	ROUGE	CIDEr
✓			0.282	0.253	0.422	0.307	0.479	0.226
✓	✓		0.297	0.264	0.419	0.310	0.479	0.228
		✓	0.351	0.323	0.430	0.325	0.495	0.263
✓		✓	0.407	0.364	0.435	0.337	0.501	0.274
✓	✓	✓	0.463	0.432	0.457	0.356	0.529	0.298

LLM, and freeze the parameters of LLM and vision tokenizer F_v. We train the prior tokenizer F_p and Q-former Q to align visual priors with text, along with the instructed vision decoder D to enhance visual localization capabilities. The fine-tuning phase equips the LLM with reasoning abilities in autonomous driving using the instructed vision decoder. To retain pre-trained LLM generalization, we employ efficient fine-tuning with LoRA [19]. The vision encoder and prior tokenizer F_p remain fixed, while the instructed vision decoder D is fully fine-tuned. Word embeddings of the LLM and Q-former are also trainable.

5 Experiments

We benchmark various baseline models and present our method on Reason2Drive dataset. Sec. 5.1 covers implementation details. We assess reasoning performance using our proposed metric in Sec. 5.2, perform ablation studies in Sec. 5.3 and provide more ablations and qualitative results in the appendix.

5.1 Experimental setting

Our main experiments are carried out on the complete Reason2Drive benchmark. The dataset is collected from three different source datasets: nuScenes [2], Waymo [37], and ONCE [28]. It is divided into training and validation sets based on segments, with 70% allocated to the training set and 30% to the validation

Table 5: The quality of predicted visual elements.

Predictions	Metric	MiniGPT-4	Kosmos-2 [33]	Ours
Bounding box	Accuracy ↑	0.723	0.745	**0.806**
Trajectory	ADE ↓	2.334	2.563	**1.875**

set, ensuring no overlap in scenes between them. The validation set also contains all the declared tasks and has been augmented. The input consists of 5 frames of cropped images with a size of 224×224 pixels. During training, we leverage the AdamW [26] optimizer with a weight decay of 0.01. We adopt a cosine learning rate decay scheduler with a max value of 3e-4 and a linear warm-up for the first 1000 iterations. The weight of perception loss λ_p is set to 1.0. The normalization parameters τ and β are selected to be 15 and 10 after empirical practice. Our models are trained with a batch size of 8 on 8 V100 GPUs. The baseline models are fine-tuned following the official tuning strategy, with the inputs consistent with our approach. Specifically, the perception priors are also provided as textual inputs to baseline models for fair comparison.

Table 6: Ablations on visual input and perception priors.

Visual features		Perceptual priors		ADRScore	ADRScore-S
image-level	video-level	region-level	positional		
✓				0.414	0.379
	✓			0.431	0.394
	✓	✓		0.447	0.418
	✓	✓	✓	0.463	0.432

Table 7: Ablations on different settings of instructed vision decoder.

Pre-train	text embedding	MLP	ADRScore	ADRScore-S
			0.387	0.361
✓			0.421	0.396
✓	✓		0.455	0.425
✓	✓	✓	0.463	0.432

5.2 Reasoning performance evaluation

As demonstrated in Tab. 3, we evaluate both ADRScore and traditional caption-based performance of different models on our benchmark. It is worth noting that our method outperforms others comprehensively in all metrics. We also observe that, despite there is a correlation between ADRScore and traditional metrics, while on the other hand, the performance gap is more pronounced in our metrics, specially revealed by ADRScore-S. The results further substantiate reasoning ambiguities in traditional metrics, constraining the differentiation in benchmarking model performance. Specifically, models with varying reasoning capabilities exhibit minimal disparities when evaluated using traditional metrics.

5.3 Ablation study

Task contributions. To investigate the synergies between different tasks, we separate and evaluate various types of tasks independently. As shown in Tab. 4, we train our model on different combinations of tasks. Most notably, training on reasoning tasks plays the most important role, indicating the necessity of relevant reasoning data for instructional tuning. Based on reasoning tasks, perception and prediction tasks additionally enhance the models for visual reasoning, showing specific improvements of 4.1% and 6.8%, respectively.

Quality of predicted visual elements. To validate the quality of predicted perceptual elements, we also conduct experiments to evaluate the bounding

Table 8: Evaluation of control signals. B: B@4. M: METHOR.

Method	LLM	B ↑	M ↑	RMSE	
				Speed ↓	Steer ↓
Directly fine-tuned with control signals:					
InstructBLIP	Vicuna-7B	0.166	0.201	3.743	5.926
Additionally pre-trained on Reason2Drive:					
InstructBLIP	Vicuna-7B	0.192	0.237	3.086	5.151
Ours	Vicuna-7B	**0.213**	**0.269**	**2.842**	**4.866**

boxes and trajectories separately, as shown in Tab. 5. To ensure a fair comparison, we implement Kosmos-2 [33] due to its advantageous grounding capabilities. The model is fine-tuned on our dataset following the official strategy. Experimental results confirm the high quality of our predicted visual elements, with particular emphasis on the accuracy of the trajectories.

The effects of tokenizers. To verify the effectiveness of the tokenizers, we conduct ablation studies to pinpoint where the improvements come from in Tab. 6. Visual features from single frame to multi-frame bring 1.5% improvement in ADRScore-S. Perceptual priors, *i.e.*, region-level features and positional embeddings bring 2.4% and 1.4% development.

The effects of instructed vision decoder. To verify the efficiency of our instructed vision decoder, we conduct an ablation study to compare it with other methods. As demonstrated in Tab. 7, pre-training and textual embedding bring the major contribution (3.5% and 2.9% in ADRScore-S).

Downstream tasks. We are also interested in understanding how our benchmark will contribute to downstream tasks, such as predicting control signals. Following the approach in ADriver-I [22], we generate control signals on nuScenes and fine-tune InstructBLIP [10] directly for planning signal prediction. To ablate the influence of our dataset, we also pre-train models on Reason2Drive before fine-tuning with control signals. The results presented in Tab. 8 demonstrate the supportive effect of Reason2Drive on downstream planning tasks.

6 Conclusion

In summary, Large Vision-Language Models (VLMs) have sparked interest in autonomous driving for their advanced reasoning capabilities. However, the absence of datasets explaining decision-making processes hinders progress. To tackle this, we introduce Reason2Drive benchmark, comprising 600K+ video-text pairs for interpretable reasoning in complex driving scenarios. It outperforms existing datasets in scale, sources and task diversities. We also propose a novel evaluation protocol for chain-based reasoning, addressing existing semantic ambiguities. To uncover insights into their reasoning abilities, our work evaluates various VLMs and proposes an efficient method to boost the ability of models to utilize object-level perceptual elements in both the encoder and decoder. We expect our work could propel further advancements in interpretable reasoning for autonomous systems.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (Grant No. 62106050 and 62376060), Natural Science Foundation of Shanghai (Grant No. 22ZR1407500), USyd-Fudan BISA Flagship Research Program and Lingang Laboratory (Grant No. LG-QS-202202-07).

References

1. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: ACL workshop (2005)
2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR (2020)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV (2020)
4. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint (2023)
5. Chen, L., Wu, P., Chitta, K., Jaeger, B., Geiger, A., Li, H.: End-to-end autonomous driving: Challenges and frontiers. arXiv preprint (2023)
6. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See https://vicuna. lmsys. org (accessed 14 April 2023) (2023)
7. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. arXiv preprint (2022)
8. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint (2022)
9. Contributors, D.: Drivelm: Drive on language. `https://github.com/OpenDriveLab/DriveLM` (2023)
10. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B.A., Fung, P., Hoi, S.C.H.: Instructblip: Towards general-purpose vision-language models with instruction tuning. arXiv preprint (2023)
11. Deruyttere, T., Grujicic, D., Blaschko, M.B., Moens, M.F.: Talk2car: Predicting physical trajectories for natural language commands. Ieee Access (2022)
12. Dewangan, V., Choudhary, T., Chandhok, S., Priyadarshan, S., Jain, A., Singh, A.K., Srivastava, S., Jatavallabhula, K.M., Krishna, K.M.: Talk2bev: Language-enhanced bird's-eye view maps for autonomous driving. arXiv preprint (2023)
13. Ding, X., Han, J., Xu, H., Zhang, W., Li, X.: Hilm-d: Towards high-resolution understanding in multimodal large language models for autonomous driving. arXiv preprint (2023)
14. Dollár, K., Girshick, R.: Mask r-cnn. In: ICCV (2017)
15. Feng, Q., Ablavsky, V., Sclaroff, S.: Cityflow-nl: Tracking and retrieval of vehicles at city scale by natural language descriptions. arXiv preprint (2021)
16. Fu, D., Li, X., Wen, L., Dou, M., Cai, P., Shi, B., Qiao, Y.: Drive like a human: Rethinking autonomous driving with large language models. arXiv preprint (2023)

17. Golovneva, O., Chen, M., Poff, S., Corredor, M., Zettlemoyer, L., Fazel-Zarandi, M., Celikyilmaz, A.: Roscoe: A suite of metrics for scoring step-by-step reasoning. arXiv preprint (2022)
18. Han, J., Zhang, R., Shao, W., Gao, P., Xu, P., Xiao, H., Zhang, K., Liu, C., Wen, S., Guo, Z., et al.: Imagebind-llm: Multi-modality instruction tuning. arXiv preprint (2023)
19. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint (2021)
20. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., et al.: Planning-oriented autonomous driving. In: CVPR (2023)
21. Jain, K., Chhangani, V., Tiwari, A., Krishna, K.M., Gandhi, V.: Ground then navigate: Language-guided navigation in dynamic scenes. In: ICRA (2023)
22. Jia, F., Mao, W., Liu, Y., Zhao, Y., Wen, Y., Zhang, C., Zhang, X., Wang, T.: Adriver-i: A general world model for autonomous driving. arXiv preprint (2023)
23. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. arXiv preprint (2023)
24. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint (2023)
25. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NIPS (2023)
26. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2017)
27. Malla, S., Choi, C., Dwivedi, I., Choi, J.H., Li, J.: Drama: Joint risk localization and captioning in driving. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2023)
28. Mao, J., Niu, M., Jiang, C., Liang, H., Chen, J., Liang, X., Li, Y., Ye, C., Zhang, W., Li, Z., et al.: One million scenes for autonomous driving: Once dataset. arXiv preprint (2021)
29. Mao, J., Qian, Y., Zhao, H., Wang, Y.: Gpt-driver: Learning to drive with gpt. arXiv preprint (2023)
30. OpenAI: Gpt-4: A large-scale transformer-based language model (2023), `https://www.openai.com/research/gpt-4`, `https://www.openai.com/research/gpt-4`
31. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL (2002)
32. Peng, B., Li, C., He, P., Galley, M., Gao, J.: Instruction tuning with gpt-4. arXiv preprint (2023)
33. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint (2023)
34. Pi, R., Gao, J., Diao, S., Pan, R., Dong, H., Zhang, J., Yao, L., Han, J., Xu, H., Zhang, L.K.T.: Detgpt: Detect what you need via reasoning. arXiv preprint (2023)
35. Qian, T., Chen, J., Zhuo, L., Jiao, Y., Jiang, Y.G.: Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. arXiv preprint (2023)
36. Sachdeva, E., Agarwal, N., Chundi, S., Roelofs, S., Li, J., Dariush, B., Choi, C., Kochenderfer, M.: Rank2tell: A multimodal driving dataset for joint importance ranking and reasoning. arXiv preprint (2023)
37. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: CVPR (2020)
38. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint (2023)

39. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: CVPR (2015)
40. Wu, D., Han, W., Wang, T., Dong, X., Zhang, X., Shen, J.: Referring multi-object tracking. In: CVPR (2023)
41. Wu, D., Han, W., Wang, T., Liu, Y., Zhang, X., Shen, J.: Language prompt for autonomous driving. arXiv preprint (2023)
42. Xu, Z., Zhang, Y., Xie, E., Zhao, Z., Guo, Y., Wong, K.K., Li, Z., Zhao, H.: Drivegpt4: Interpretable end-to-end autonomous driving via large language model. arXiv preprint (2023)
43. Yu, P., Wang, T., Golovneva, O., Alkhamissy, B., Ghosh, G., Diab, M., Celikyilmaz, A.: Alert: Adapting language models to reasoning tasks. arXiv preprint (2022)
44. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint (2023)
45. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint (2022)
46. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint (2023)