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] 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 semantic 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. The code and dataset will be released.

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 [19] have been developed to derive control signals directly from sensor inputs, treating the system as a black 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 [33, 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.

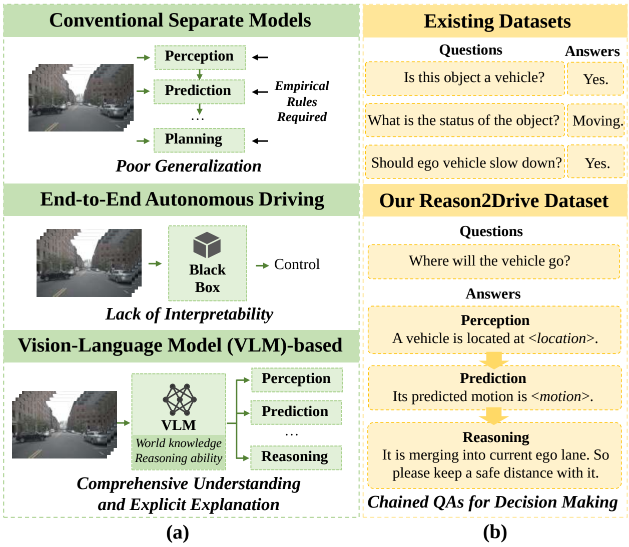

Conventional Separate Models
Perception
Prediction
...
Planning
Empirical Rules Required
Poor Generalization
End-to-End Autonomous Driving
Black Box
Control
Lack of Interpretability
Vision-Language Model (VLM)-based
VLM
World knowledge
Reasoning ability
Perception
Prediction
...
Reasoning
Comprehensive Understanding and Explicit Explanation
(a)
Existing Datasets
Questions
Answers
Is this object a vehicle?
Yes.
What is the status of the object?
Moving.
Should ego vehicle slow down?
Yes.
Our Reason2Drive Dataset
Questions
Where will the vehicle go?
Answers
Perception
A vehicle is located at <location>.
Prediction
Its predicted motion is <motion>.
Reasoning
It is merging into current ego lane. So please keep a safe distance with it.
Chained QAs for Decision Making
(b)

Figure 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 [10, 33, 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 [36], and ONCE [26], utilizing an extensible annotation schema. Specifically, we extract object metadata, structure it into JSON format, and integrate it into pre-defined templates to create paired data for VLMs at both object and scenario levels. To enhance diversity, GPT-4 and manual annotations 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 semantic ambiguity of traditional caption-based metrics like BLEU [29] and CIDEr [39]. For example, sentences with contrasting meanings such as “It will turn left” and “It will turn right” could yield high scores in BLEU, which is especially problematic in the context of autonomous driving. To address this issue, we propose a new aggregated evaluation metric specifically designed to measure chain-based reasoning performance in autonomous systems, which aims to resolve the semantic ambiguities associated with current metrics.

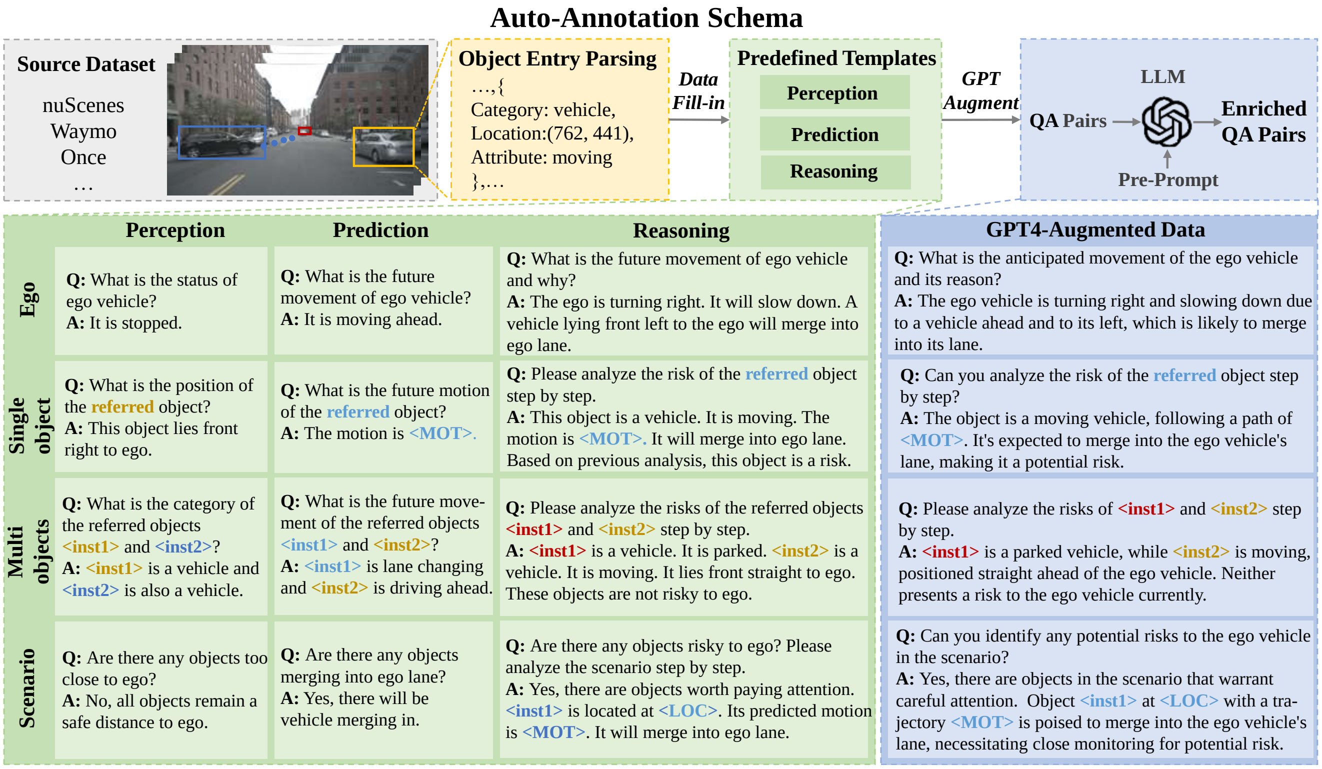
Utilizing the proposed benchmark, we undertake experiments to assess various existing VLMs, thereby unveiling valuable insights into their reasoning capabilities. We find that most methods struggle to effectively leverage perceptual priors, resulting in subpar reasoning performance. Additionally, constrained 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.

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 to assess chain-based reasoning performance in autonomous driving, effectively addressing the semantic 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 challenge posed by inefficient priors feature extraction and inaccurate perceptual predictions, we introduce an efficient approach for integrating these into VLMs, resulting in a substantial improvement in reasoning accuracy. 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 ([5, 6, 28, 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 [22] aligns visual and language features by utilizing a learnable Q-former. LLaVA [23] and MiniGPT-4 [48] initially align image-text features and then proceed with instruction tuning. Additionally, Video-LLaMA [46] and ImageBind-LLM [16] 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 [31] and Shikra [4] perform object detection based on instructions and also accomplish grounded visual question answering. DetGPT [32] connects a fixed multimodal LLM with a customizable detector based on user instructions. LISA [21] 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.

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 [12, 14, 27, 42, 45]. Simultaneously, substantial works are in progress to create datasets and models tailored to various tasks. Talk2Car [10] proposes the first object referral dataset for grounding com-



	Perception	Prediction	Reasoning	GPT4-Augmented Data
Ego	**Q:** What is the status of ego vehicle? **A:** It is stopped.	**Q:** What is the future movement of ego vehicle? **A:** It is moving ahead.	**Q:** What is the future movement of ego vehicle and why? **A:** The ego is turning right. It will slow down. A vehicle lying front left to the ego will merge into ego lane.	**Q:** What is the anticipated movement of the ego vehicle and its reason? **A:** The ego vehicle is turning right and slowing down due to a vehicle ahead and to its left, which is likely to merge into its lane.
Single object	**Q:** What is the position of the **referred** object? **A:** This object lies front right to ego.	**Q:** What is the future motion of the **referred** object? **A:** The motion is <MOT>.	**Q:** Please analyze the risk of the **referred** object step by step. **A:** This object is a vehicle. It is moving. The motion is <MOT>. It will merge into ego lane. Based on previous analysis, this object is a risk.	**Q:** Can you analyze the risk of the **referred** object step by step? **A:** The object is a moving vehicle, following a path of <MOT>. It's expected to merge into the ego vehicle's lane, making it a potential risk.
Multi objects	**Q:** What is the category of the referred objects <inst1> and <inst2>? **A:** <inst1> is a vehicle and <inst2> is also a vehicle.	**Q:** What is the future movement of the referred objects <inst1> and <inst2>? **A:** <inst1> is lane changing and <inst2> is driving ahead.	**Q:** Please analyze the risks of the referred objects <inst1> and <inst2> step by step. **A:** <inst1> is a vehicle. It is parked. <inst2> is a vehicle. It is moving. It lies front straight to ego. These objects are not risky to ego.	**Q:** Please analyze the risks of <inst1> and <inst2> step by step. **A:** <inst1> is a parked vehicle, while <inst2> is moving, positioned straight ahead of the ego vehicle. Neither presents a risk to the ego vehicle currently.
Scenario	**Q:** Are there any objects too close to ego? **A:** No, all objects remain a safe distance to ego.	**Q:** Are there any objects merging into ego lane? **A:** Yes, there will be vehicle merging in.	**Q:** Are there any objects risky to ego? Please analyze the scenario step by step. **A:** Yes, there are objects worth paying attention. <inst1> is located at <LOC>. Its predicted motion is <MOT>. It will merge into ego lane.	**Q:** Can you identify any potential risks to the ego vehicle in the scenario? **A:** Yes, there are objects in the scenario that warrant careful attention. Object <inst1> at <LOC> with a trajectory <MOT> is poised to merge into the ego vehicle's lane, necessitating close monitoring for potential risk.

Figure 2. Schematic illustration 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.

mands for self-driving cars in free natural language into the visual context. But it exclusively contains information about visible objects. While DRAMA [25] outlines the overall scene risk, it lacks precise perception annotation. NuPrompt [41] and Refer-KITTI [40] offer language prompt sets for driving scenes but primarily concentrate on multi-object tracking tasks. NuScenesQA [33] and DriveLM [8] 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 semantic 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 object metadatas into JSON-structured entries. Each object entry contains various details pertaining to its driving actions, including location, category, attributes and more. Afterwards, these extracted entries are filled into predefined templates, which are divided into different tasks (i.e., perception, prediction and reasoning) at both object-level and scenario-level. Subsequently, GPT-4 and manual annotations are involved for verification and enrichment purposes.

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 focused 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, 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 and meaningful.

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 statistic the distribution of our dataset in Fig. 4. The benchmark exhibits a balanced distribution, with multi-object tasks constituting the majority. Additionally, perception, prediction and reasoning questions are distributed as 39%, 34%, and 27%, respectively. More details are provided in the appendix.

3.3. Benchmark protocol

It is worth noting that previous works [11, 25, 33] simply utilize metric scores widely used in the NLP community, including BLEU [29], CIDEr [39] and METEOR [1]. However, these metrics primarily measure word-level performance and do not account for semantic meaning, which may lead to unexpected evaluation results. To address the semantic ambiguities, inspired by [44] and [15], we develop the evaluation protocol to measure the correctness of the reasoning chains.

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

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

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

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. 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

Dataset	Description	Source Datasets
Talk2Car [10] 🟧	Object referral	nuScenes
CityFlow-NL [13] 🩷	Tracking & retrieval	CityFlow
CARLA-NAV [20] 🟨	Segmentation & prediction	CARLA Simulator
NuPrompt [41] ⬛	Multi-object tracking	nuScenes
NuScenes-QA [33] 🟦	Perception	nuScenes
Refer-KITTI [40] 🟩	Multi-object tracking	KITTI
Talk2BEV [11] ⬛	Visual understanding	nuScenes
DRAMA [25] 🟪	Risk localization	self-collected
Rank2Tell [35] 🟦	Risk localization & ranking	self-collected
Reason2Drive🟥	Perception Prediction Reasoning	nuScenes Waywo ONCE

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

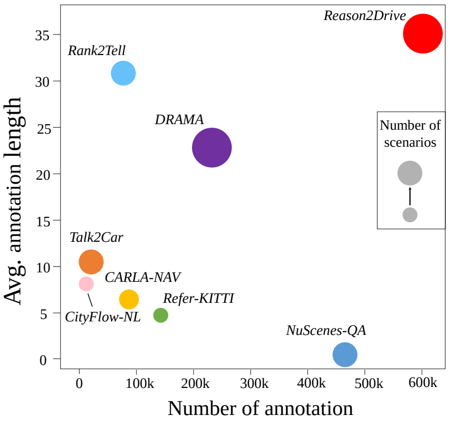


Figure 3. Data quality comparison. Reason2Drive is larger in scale, richer in data content, and more diverse in scenarios.

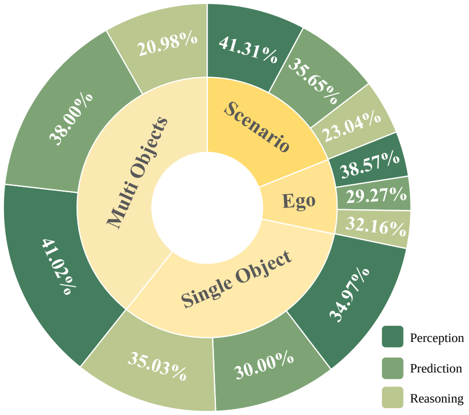


Figure 4. Statistical distribution of different tasks in Reason2Drive dataset, which illustrates the equilibrium of our proposed dataset.

reasoning alignment between them:

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

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.

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

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:

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

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

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

Strict reason. To further adapt to the realistic driving process, we promote the above metric to the situation with visual elements. Specifically, when the hypothesis step h_i and reference step r_k contains visual elements, *i.e.*, the locations and motions predicted for further reasoning, the similarity score becomes:

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

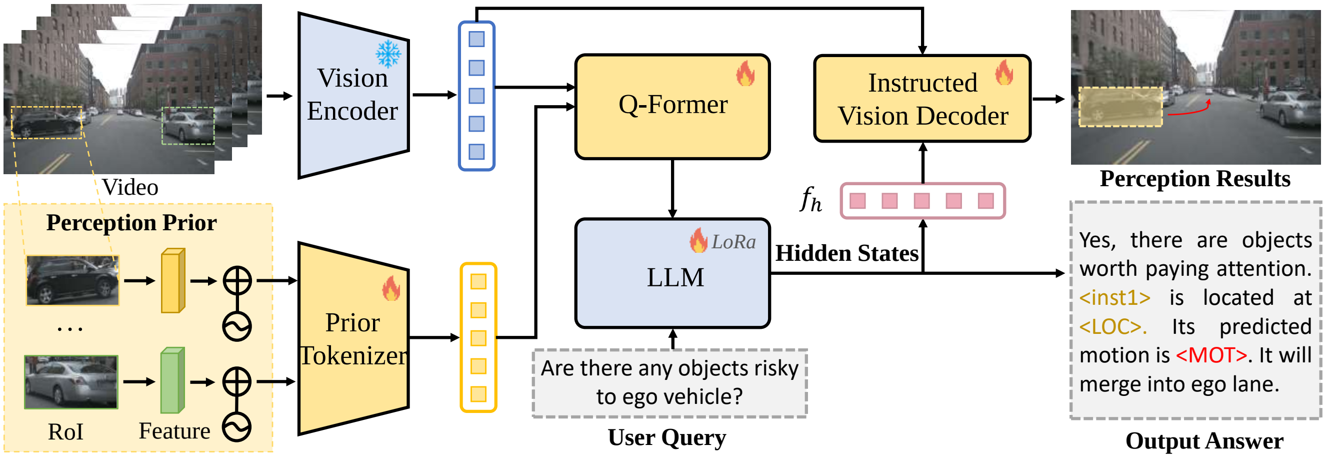
where $M(\cdot)$ measures the mean square error between two perceptual elements. And we normalize it to $[0, 1]$ to match the distribution of semantic-level similarity. The promoted strict reason metric is designed to more precisely assess the reasoning responses containing perceptual elements.

4. Method

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



Are there any objects risky to ego vehicle?

Yes, there are objects worth paying attention. <inst1> is located at <LOC>. Its predicted motion is <MOT>. It will merge into ego lane.

Figure 5. The pipeline of our proposed 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 result image respectively represent the visualization of `<LOC>` and `<MOT>`.

indispensable in autonomous driving scenarios. The limitation is primarily due to (i) the lack of a targeted tokenizer and (ii) decoder solely composed of language model, resulting in subpar reasoning performance. To address this challenge, as illustrated in Fig. 5, we introduce a straightforward yet effective framework that enhances existing VLMs with two new components: a prior tokenizer and an instructed vision decoder. These components aim to strengthen the capabilities of the model to utilize object-level perceptual elements in both the process of extracting visual priors and generating perceptual predictions.

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 [22] 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 [17] 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 [22] 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 [10, 14] 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 [21] have demon-

Methods	LLM	Reasoning metric		Captioning metric			
		Strict Reason	Reason	B@4	METEOR	ROUGE	CIDEr
Blip-2 [22]	OPT-2.7B [47]	0.162	0.296	0.361	0.249	0.443	0.174
	FlanT5-XL [7]	0.171	0.310	0.368	0.256	0.451	0.187
InstructBLIP [9]	FlanT5-XL	0.187	0.329	0.376	0.269	0.462	0.196
	Vicuna-7B [30]	0.214	0.351	0.408	0.294	0.484	0.211
MiniGPT-4 [48]	Vicuna-7B	0.203	0.338	0.396	0.286	0.475	0.219
Ours	FlanT5-XL	0.420	0.457	0.451	0.349	0.520	0.292
	Vicuna-7B	**0.432**	**0.463**	**0.457**	**0.356**	**0.529**	**0.298**

Table 2. Results of different models on the Reason2Drive validation set. We evaluate the reasoning metrics as well as captioning metrics.

Perception	Prediction	Reasoning	Strict Reason	Reason
✓			0.253	0.282
✓	✓		0.264	0.297
		✓	0.323	0.351
✓		✓	0.364	0.407
✓	✓	✓	0.432	0.463

Table 3. Ablations on different combinations of training tasks.

strated 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 [21], 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 alignments [3] and task-specific heads designed to generate object locations and motions independently.

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 balanced 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 [9], including the pre-trained vision encoder, Q-former and 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 [18]. 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 qualitative results in the appendix.

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

Table 4. Ablations on visual input and perception priors.

Pre-train	text embedding	MLP	Strict Reason	Reason
			0.361	0.387
✓			0.396	0.421
✓	✓		0.425	0.455
✓	✓	✓	0.432	0.463

Table 5. Ablations on different settings of instructed vision decoder.

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 [36], and ONCE [26]. It is divided into training and validation sets based on segments, with 70% allocated to the training set and 30% to the validation set, ensuring no overlap in scenes between them. The input consists of 5 frames of cropped images with a size of 224×224 pixels. During training, we leverage the AdamW [24] 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 for 10 epochs with a batch size of 8 on 8 V100 GPUs.

5.2. Reasoning results

As demonstrated in Tab. 2, we not only evaluate the reasoning scores of different models on our benchmark but also assess their performance using traditional caption-based evaluation metrics. It is worth noting that our method outperforms others comprehensively, both in terms of reasoning scores and traditional metrics. In detail, the performance of the LLM plays a significant role. We observe that, on the one hand, there is a correlation between our reasoning scores and traditional metrics, while on the other hand, the performance gap is more pronounced in our metrics.

Methods	LLM	Strict Reason	Reason	GPT-3.5	GPT-4
Blip-2	OPT-2.7B	0.332	0.450	0.479	0.458
InstructBLIP	FlanT5-XL	0.377	0.489	0.532	0.501
MiniGPT-4	Vicuna-7B	0.352	0.469	0.519	0.467
Ours	Vicuna-7B	**0.561**	**0.593**	**0.643**	**0.628**

Table 6. Evaluation results given by prompted ChatGPT.

Method	LLM	Training	Testing	
			N	W + O
Blip-2	OPT-2.7B	N	0.205	0.116
		W + O	0.197	0.121
InstructBLIP	FlanT5-XL	N	0.238	0.155
		W + O	0.255	0.149
MiniGPT-4	Vicuna-7B	N	0.257	0.172
		W + O	0.263	0.168
Ours	Vicuna-7B	N	**0.443**	**0.397**
		W + O	**0.428**	**0.385**

Table 7. Generalization ability when transferred to different sources of datasets. Strict reason metric is reported.

5.3. Ablation study

Task contributions. To investigate the synergies between different tasks, we evaluate the tasks independently. As shown in Tab. 3, training on reasoning tasks contributed the most. Meanwhile, the perception tasks and prediction tasks contribute 4.1% and 6.8% respectively (Row 3, 4 and 5).

The effects of tokenizers. To verify the effectiveness of the tokenizers, we conduct ablation studies to pinpoint where the improvements come from (Tab. 4). Visual features from single frame to multi-frame bring 1.5% improvement. 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. 5, pre-training and textual embedding bring the major contribution (3.5% and 2.9% in strict reason).

Evaluated by GPT-4. To validate the rationality of our reasoning scores, following [14], we employ GPT-4 to validate the generated answers in Tab. 6. We can draw the conclusion that our method still achieves superior performance, which also indicates the rationality of our proposed metric.

Generalization. To validate the method's generalization, we trained on the Reason2Drive benchmark with only the nuScenes dataset and tested on Waymo and ONCE in Tab. 7. We split the Reason2Drive benchmark into two sets, nuScenes (noted as N) and Waymo + ONCE (noted as W + O). Compared with others, our method suffers limited performance drops (4.6% and 4.3%).

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. Code and dataset will be released.

References

[1] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL workshop*, 2005. 4

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 2, 8

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 7

[4] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint*, 2023. 2

[5] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *https://vicuna.lmsys.org*, 2023. 2, 11

[6] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint*, 2022. 2

[7] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint*, 2022. 7, 11

[8] DriveLM Contributors. Drivelm: Drive on language. https://github.com/OpenDriveLab/DriveLM, 2023. 3

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint*, 2023. 7, 11

[10] Thierry Deruyttere, Dusan Grujicic, Matthew B Blaschko, and Marie-Francine Moens. Talk2car: Predicting physical trajectories for natural language commands. *Ieee Access*, 2022. 1, 2, 5, 6

[11] Vikrant Dewangan, Tushar Choudhary, Shivam Chandhok, Shubham Priyadarshan, Anushka Jain, Arun K Singh, Siddharth Srivastava, Krishna Murthy Jatavallabhula, and K Madhava Krishna. Talk2bev: Language-enhanced bird's-eye view maps for autonomous driving. *arXiv preprint*, 2023. 4, 5

[12] Xinpeng Ding, Jianhua Han, Hang Xu, Wei Zhang, and Xiaomeng Li. Hilm-d: Towards high-resolution understanding in multimodal large language models for autonomous driving. *arXiv preprint*, 2023. 2

[13] Qi Feng, Vitaly Ablavsky, and Stan Sclaroff. Cityflow-nl: Tracking and retrieval of vehicles at city scale by natural language descriptions. *arXiv preprint*, 2021. 5

[14] Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. Drive like a human: Rethinking autonomous driving with large language models. *arXiv preprint*, 2023. 2, 6, 8

[15] Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. Roscoe: A suite of metrics for scoring step-by-step reasoning. *arXiv preprint*, 2022. 4

[16] Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint*, 2023. 2

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 6

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint*, 2021. 7

[19] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *CVPR*, 2023. 1

[20] Kanishk Jain, Varun Chhangani, Amogh Tiwari, K Madhava Krishna, and Vineet Gandhi. Ground then navigate: Language-guided navigation in dynamic scenes. In *ICRA*, 2023. 5

[21] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmenta-

tion via large language model. *arXiv preprint*, 2023. 2, 6, 7

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint*, 2023. 2, 6, 7

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint*, 2017. 8

[25] Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. Drama: Joint risk localization and captioning in driving. In *WACV*, 2023. 3, 4, 5

[26] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, et al. One million scenes for autonomous driving: Once dataset. *arXiv preprint*, 2021. 2, 8

[27] Jiageng Mao, Yuxi Qian, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. *arXiv preprint*, 2023. 2

[28] OpenAI. Gpt-4: A large-scale transformer-based language model, 2023. https://www.openai.com/research/gpt-4. 2

[29] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 2, 4

[30] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint*, 2023. 7

[31] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint*, 2023. 2

[32] Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, Lingpeng Kong, and Tong Zhang. Detgpt: Detect what you need via reasoning. *arXiv preprint*, 2023. 2

[33] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. *arXiv preprint*, 2023. 1, 3, 4, 5

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 11

[35] Enna Sachdeva, Nakul Agarwal, Suhas Chundi, Sean Roelofs, Jiachen Li, Behzad Dariush, Chiho Choi, and Mykel Kochenderfer. Rank2tell: A multimodal driving dataset for joint importance ranking and reasoning. *arXiv preprint*, 2023. 5

[36] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 2, 8

[37] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint*, 2023. 11

[38] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint*, 2023. 2, 11

[39] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 2, 4

[40] Dongming Wu, Wencheng Han, Tiancai Wang, Xingping Dong, Xiangyu Zhang, and Jianbing Shen. Referring multi-object tracking. In *CVPR*, 2023. 1, 3, 5

[41] Dongming Wu, Wencheng Han, Tiancai Wang, Yingfei Liu, Xiangyu Zhang, and Jianbing Shen. Language prompt for autonomous driving. *arXiv preprint*, 2023. 1, 3, 5

[42] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kenneth KY Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *arXiv preprint*, 2023. 2

[43] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint*, 2020. 11

[44] Ping Yu, Tianlu Wang, Olga Golovneva, Badr Alkhamissy, Gargi Ghosh, Mona Diab, and Asli Celikyilmaz. Alert: Adapting language models to reasoning tasks. *arXiv preprint*, 2022. 4

[45] Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. Contextual object detection with multimodal large language models. *arXiv preprint*, 2023. 2

[46] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint*, 2023. 2

[47] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint*, 2022. 7

[48] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint*, 2023. 2, 7

A. Appendix

A.1. More statistical analysis of Reason2Drive

In this section, we present more dataset details. As demonstrated in Tab. 8, we split the dataset according to the task and target. The benchmark exhibits a balanced distribution. Specifically, multi-object questions constitute the majority, followed by single-object and scenario-level questions, which 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.

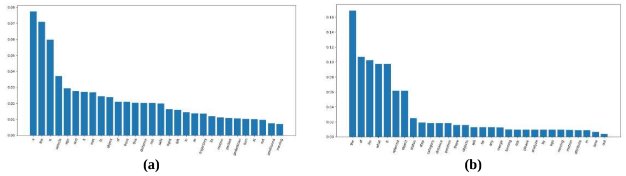


Figure 6. Words distributions in (a) questions and (b) annotated answers.

We also statistic the distribution of the words, as is illustrated in Fig. 6. From the words distribution, we can observe that Reason2Drive has a large range of words that describe perceptions, predictions and reasoning tasks, like “moving”, “distance”, and “risk”.

A.2. More implementation details

Architecture. For the frozen visual encoder, we employ ViT-G/14 from EVA-CLIP [37] in the main paper, which is a state-of-the-art pre-trained vision transformer models. We remove the last layer of the ViT and uses the second last layers' output features.

For the language model, we explore two types of LLMs: encoder-decoder-based LLMs and decoder-based LLMs. For encoder-decoder-based LLMs, we employ FlanT5-XL [7], which is an instruction-tuned model based on the encoder-decoder Transformer T5 [43]. For decoder-based LLMs, we select Vicuna [5], a recently released decoder-only Transformer instruction-tuned from LLaMA [38].

Training loss. Our model is trained with a language modelling loss $\mathcal{L}_{txt}$, where the task of the frozen LLM is to generate text conditioned on the extracted modality features of the Q-former. Furthermore, we employ an auxiliary perception loss $\mathcal{L}_{per}$ to enhance the perceptual capability. Specifically, a linear combination of a binary cross-entropy loss for classification and a regression loss is defined:

$$\mathcal{L}_{per}(P, \hat{P}) = -\sum_{i=1}^{N} log\hat{P}_{c,i} + \lambda_{reg} \sum_{i=1}^{N} \mathcal{L}_{reg}(P_{b,i}, \hat{P}_{b,i}), \tag{14}$$

where $\hat{P}_{c,i}$ and $\hat{P}_{b,i}$ are predicted classification and regression results of $\hat{P}$. Loss function $\mathcal{L}_{reg}$ is employed by a MSE loss. In practice, we select λ_{reg} to be 0.25 as the balance term as a common setting in object detection tasks.

A.3. Ablation of visual encoders

We ablate the effects of employed visual encoders in Tab. 9. For comparison, we explore two types of visual encoders: ViT-L/14 in CLIP [34] and ViT-G/14 in EVA-CLIP [37]. We can draw the conclusion that the performance of visual encoder inevitably influences the VLMs especially in strict reason metric.

Task Target	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 8. The statistics of different tasks in Reason2Drive dataset.

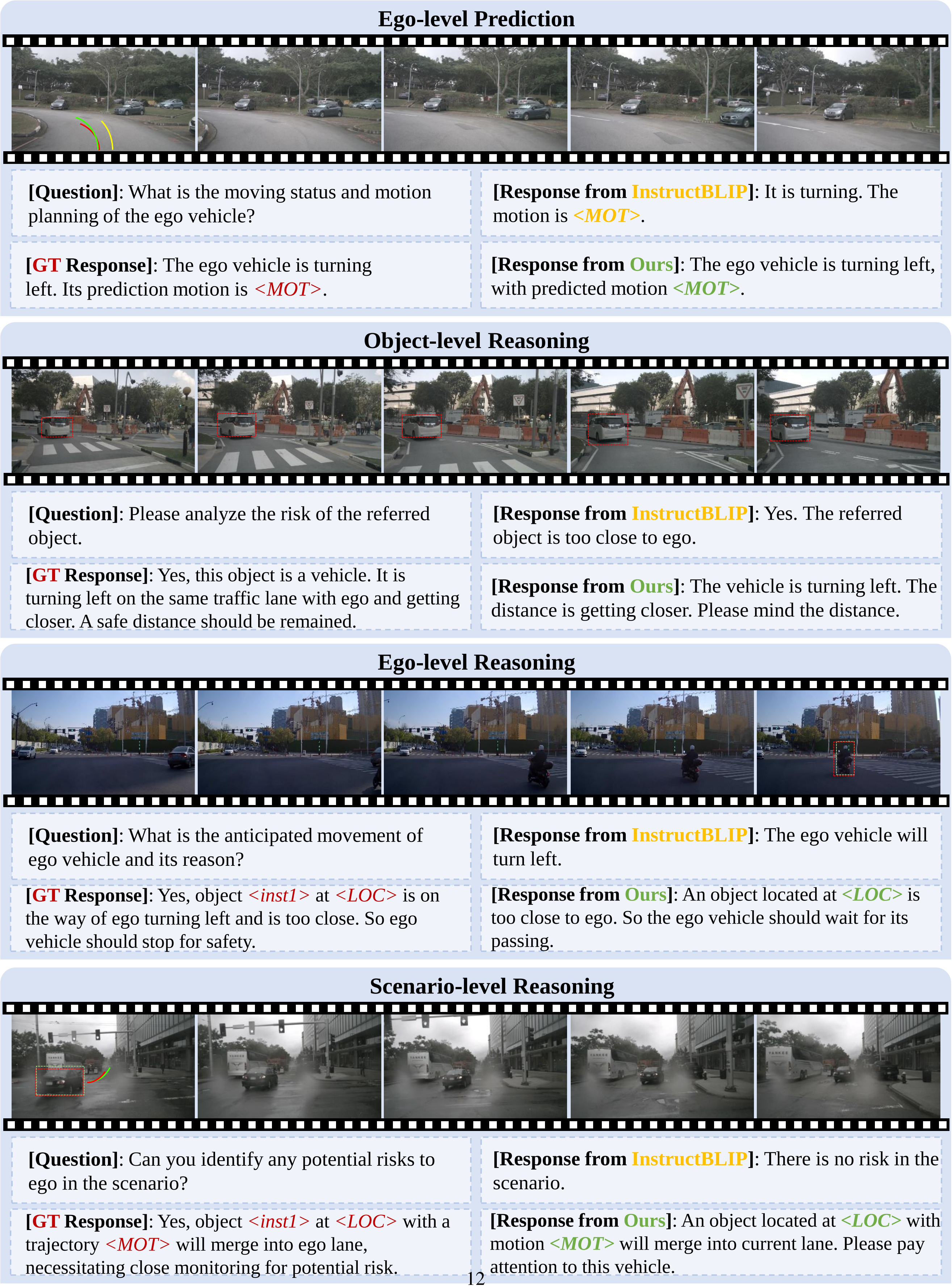
Method	Visual encoder	Strict Reason	Reason
Blip-2	ViT-L/14 [34]	0.155	0.294
	ViT-G/14 [37]	0.171	0.310
InstructBLIP	ViT-L/14	0.187	0.327
	ViT-G/14	0.214	0.351
Ours	ViT-L/14	0.397	0.435
	ViT-G/14	**0.432**	**0.463**

Table 9. Ablations on visual encoders.

A.4. Qualitative examples

Successful cases. In Fig. 7, we visualize some of the successful cases in our Reason2Drive validation set. In general, our method behaves better than InstructBLIP [9] in most scenarios. Our method performs well on the planning prediction of objects, the recognition of potential risks and reasoning steps under different levels of tasks. The qualitative results demonstrate the effectiveness of our method towards interpretable and chain-based reasoning, which has great implications for autonomous driving.

Failure cases. In Fig. 8, we show the generation failures. For some relatively complex driving scenarios, the existing methods, including ours, still make some mistakes. In the first case of ego-level prediction, the network predicted the stooped ego vehicle to be turning because the slightly movement of the ego vehicle. In the second and third cases of object-level perception and prediction, both our method and InstructBLIP misjudged the moving status of the referred object due to the relative displacement of the ego car. Besides, the VLMs seem likely to miss recognition when opposed to distant risk objects, as illustrated in the fourth case. These issues may be mitigated by targeted research to enhance the features of distance objects and the encoding of dynamic displacement of the ego vehicle in the future.



Ego-level Prediction

[Question]: What is the moving status and motion planning of the ego vehicle?

[Response from InstructBLIP]: It is turning. The motion is *<MOT>*.

[GT Response]: The ego vehicle is turning left. Its prediction motion is *<MOT>*.

[Response from Ours]: The ego vehicle is turning left, with predicted motion *<MOT>*.

Object-level Reasoning

[Question]: Please analyze the risk of the referred object.

[Response from InstructBLIP]: Yes. The referred object is too close to ego.

[GT Response]: Yes, this object is a vehicle. It is turning left on the same traffic lane with ego and getting closer. A safe distance should be remained.

[Response from Ours]: The vehicle is turning left. The distance is getting closer. Please mind the distance.

Ego-level Reasoning

[Question]: What is the anticipated movement of ego vehicle and its reason?

[Response from InstructBLIP]: The ego vehicle will turn left.

[GT Response]: Yes, object *<inst1>* at *<LOC>* is on the way of ego turning left and is too close. So ego vehicle should stop for safety.

[Response from Ours]: An object located at *<LOC>* is too close to ego. So the ego vehicle should wait for its passing.

Scenario-level Reasoning

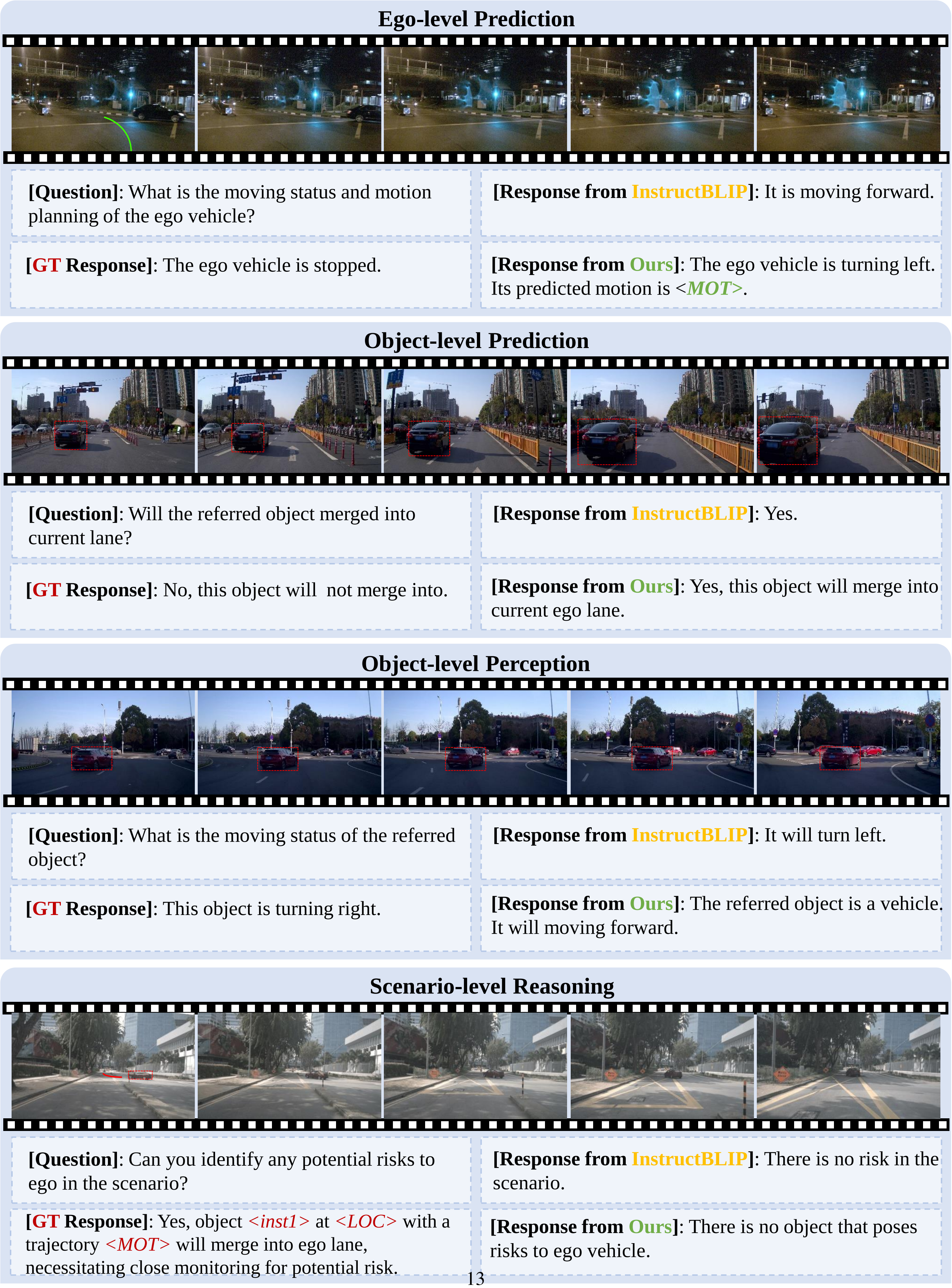
[Question]: Can you identify any potential risks to ego in the scenario?

[Response from InstructBLIP]: There is no risk in the scenario.

[GT Response]: Yes, object *<inst1>* at *<LOC>* with a trajectory *<MOT>* will merge into ego lane, necessitating close monitoring for potential risk.

[Response from Ours]: An object located at *<LOC>* with motion *<MOT>* will merge into current lane. Please pay attention to this vehicle.

Figure 7. Successful examples. Locations and motions are pictured in the first frame for better visualization. Ground truth in red color and prediction in green color.



Ego-level Prediction

[Question]: What is the moving status and motion planning of the ego vehicle?

[Response from InstructBLIP]: It is moving forward.

[GT Response]: The ego vehicle is stopped.

[Response from Ours]: The ego vehicle is turning left. Its predicted motion is <*MOT*>.

Object-level Prediction

[Question]: Will the referred object merged into current lane?

[Response from InstructBLIP]: Yes.

[GT Response]: No, this object will not merge into.

[Response from Ours]: Yes, this object will merge into current ego lane.

Object-level Perception

[Question]: What is the moving status of the referred object?

[Response from InstructBLIP]: It will turn left.

[GT Response]: This object is turning right.

[Response from Ours]: The referred object is a vehicle. It will moving forward.

Scenario-level Reasoning

[Question]: Can you identify any potential risks to ego in the scenario?

[Response from InstructBLIP]: There is no risk in the scenario.

[GT Response]: Yes, object *<inst1>* at *<LOC>* with a trajectory *<MOT>* will merge into ego lane, necessitating close monitoring for potential risk.

[Response from Ours]: There is no object that poses risks to ego vehicle.

Figure 8. Failure examples. Locations and motions are pictured in the first frame for better visualization. Ground truth in red color and prediction in green color.