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

A More Details of Reason2Drive

A.1 Words distribution

We count the distribution of the words, as is illustrated in Fig. 1. 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 Detailed sub-tasks in Reason2Drive

In this section, we present more dataset details about Reason2Drive. As introduced in the main paper, we divide the autonomous driving tasks to three distinct groups to acquire diversified data: perception, prediction and reasoning. In detail, we have a further breakdown of driving tasks, covering 15 sub-tasks for perception, 14 for prediction and 6 for reasoning, with specific examples provided in Tab. 1.

A.3 Prompts and human instructions

In Fig 2, we show the prompts and human instructions for generating augmented question-answer pairs. We provide system prompts for GPT of being an AI assistant designed for augmenting question-answer pairs. For each sample in the given examples, the "content" has the exemplar question-answer pairs, and the "response" refers to human-written instructions for demonstration. Finally, the real question-answer pairs are provided in the user's content.

B More Implementation Details

B.1 Architecture

For the frozen visual encoder, we employ ViT-G/14 from EVA-CLIP [6] in the main paper, which is a state-of-the-art pre-trained vision transformer models.

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

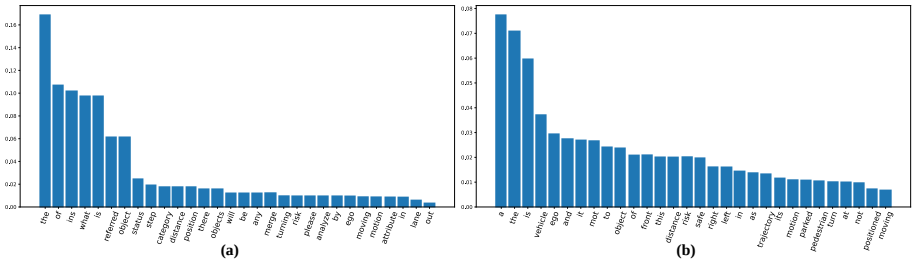

Fig. 1: Words distributions in (a) questions and (b) annotated answers.

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 [2], which is an instruction-tuned model based on the encoder-decoder Transformer T5 [8]. For decoder-based LLMs, we select Vicuna [1], a recently released decoder-only Transformer instruction-tuned from LLaMA [7].

B.2 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:

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

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.

B.3 Implementation of baseline models

The baseline models are implemented following the official implementation and fine-tuned on Reason2Drive, with inputs consistent with our approach. Specifically, perception priors are provided as textual inputs to the baseline models to ensure a fair comparison. Since baseline models lack a vision decoder, we prompt the fine-tuned baseline models to output perceptual results in textual form. These results can be identified using regular expressions for the evaluation of ADRScore-S.

```
messages=[{"role": "system", "content": "you are an AI assistant designed
for generating augmented versions of given question-answer pairs. Your task
is to make the generated question-answer pairs more diverse. Here's how
you can accomplish the task:
## INSTRUCTIONS:
- Stick to the facts described in the given question-answer pairs.
- Consider changing synonyms or changing the word orders.
- Do not change the special tokens <LOC> and <MOT> in the given
question-answer pairs.
You will also be provided with some examples:"}]
for sample in samples:
  messages.append([{"role": "user", "content": sample['content']}])
  messages.append([{"role": "assistant", "content": sample['response']}])
messages.append(["role": "user", "content": "Please augment the following
question-answer pair:\n"
f"Question: {question}\n"
f"Answer: {answer}\n"
Your response should be in the form of a Python dictionary string with keys
'Question' and 'Answer'."])
```

Fig. 2: Prompts and human instructions on augmenting question-answer pairs.

C More Ablations

C.1 Ablation of visual encoders

We ablate the effects of employed visual encoders in Tab. 2. For comparison, we explore two types of visual encoders: ViT-L/14 in CLIP [5] and ViT-G/14 in EVA-CLIP [6]. We can draw the conclusion that the performance of visual encoder inevitably influences the VLMs especially in strict reason metric.

C.2 Evaluated by GPT-4

To validate the rationality of our reasoning scores, following [4], we employ GPT-4 to validate the generated answers in Tab. 3. We can draw the conclusion that our method still achieves superior performance, which also indicates the rationality of our proposed metric. In addition, our metrics are not gpt-dependent, thus they are not affected by the migration of GPT. As a contrast, in addition to being more interpretable, our metrics are not affected by the migration of GPT, ensuring the independence and stability.

C.3 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. 4. 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. The generalization results suggest that the world knowledge of LLM helps the model generalized to the unseen scenarios. Thereby we observe that there is no significant gap between training from different sources.

D Qualitative Examples

D.1 Successful Cases

In Fig. 3, we visualize some of the successful cases in our Reason2Drive validation set. In general, our method behaves better than InstructBLIP [3] 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.

D.2 Failure Cases

In Fig. 4, 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.

References

1. 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)
2. 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)

3. 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)
4. 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)
5. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
6. Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y.: Eva-clip: Improved training techniques for clip at scale. arXiv preprint (2023)
7. 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)
8. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mt5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934 (2020)

Task	Sub-task	Target	Template
Perception	Category	Single object	What is the category of the referred object?
		Multi objects	Are any of these objects vehicles?
		Scenario	How many vehicles in the driving scenario?
	Attribute	Single object	What is the moving status of the referred object?
		Multi objects	Which of these objects is stopped?
		Scenario	Are there any objects parked in the driving scenario?
	Distance	Single object	What is the distance of the referred object towards ego?
		Multi objects	Which of these objects is closest to the ego?
		Scenario	Are there any objects too close to the ego in the driving scenario?
	Position	Single object	What is the position of the referred object?
		Multi objects	Which of these objects is located at left of the ego?
		Scenario	Are there any objects right in front of the ego in the driving scenario?
Prediction	Motion	Single object	What is the future trajectory of the referred object?
		Ego	What is the future trajectory of the ego vehicle?
	Moving strategy	Single object	What will the moving status of the referred object be in a few seconds?
		Multi objects	Which of these objects will be stopped in a few seconds?
		Ego	What will the moving status of the ego vehicle be in a few seconds?
	Turn	Single object	Which direction will the referred object turn?
		Multi objects	Which of these objects will turn left?
		Scenario	Will there be any objects turning right in the driving scenario?
	Trend	Single object	Will the referred object approach or stay away?
		Multi objects	Which of these objects will approach?
		Scenario	Will there be any objects approaching the ego vehicle?
	Merge	Single object	Will the referred object merge in/out of the ego lane?
		Multi objects	Which of these objects will merge in/out of the ego lane?
		Scenario	Will there be any objects merging in/out of the ego lane?
Reasoning	Driving strategy	Single object	What is the referred object doing and what causes it?
		Ego	What is the ego vehicle doing and what causes it?
	Risk	Single object	Is the referred object risky to the ego vehicle's normal driving?
		Scenario	Is there any risk to the ego vehicle's normal driving in the scenario?
	Control	Single object	What will the referred object do in a few seconds for safety driving and why?
		Ego	What will the ego vehicle do in a few seconds for safety driving and why?

Table 1: Details of sub-tasks and question templates.

Method	Visual encoder	ADRScore	ADRScore-S
Blip-2	ViT-L/14 [5]	0.294	0.155
	ViT-G/14 [6]	0.310	0.171
InstructBLIP	ViT-L/14	0.327	0.187
	ViT-G/14	0.351	0.214
Ours	ViT-L/14	0.435	0.397
	ViT-G/14	**0.463**	**0.432**

Table 2: Ablations on visual encoders.

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

Table 3: Evaluation results given by prompted ChatGPT.

Method	LLM	Training	Testing	
			N	W + O
Blip-2	OPT-2.7B	N	0.205	0.104
		W + O	0.183	0.121
InstructBLIP	FlanT5-XL	N	0.255	0.116
		W + O	0.212	0.155
MiniGPT-4	Vicuna-7B	N	0.263	0.130
		W + O	0.226	0.172
Ours	Vicuna-7B	N	**0.443**	**0.385**
		W + O	**0.428**	**0.397**

Table 4: Generalization ability when transferred to different sources of datasets. ADRScore-S is reported.

Fig. 3: Successful examples. Locations and motions are pictured in the first frame for better visualization. Ground truth in red color and prediction in green color.

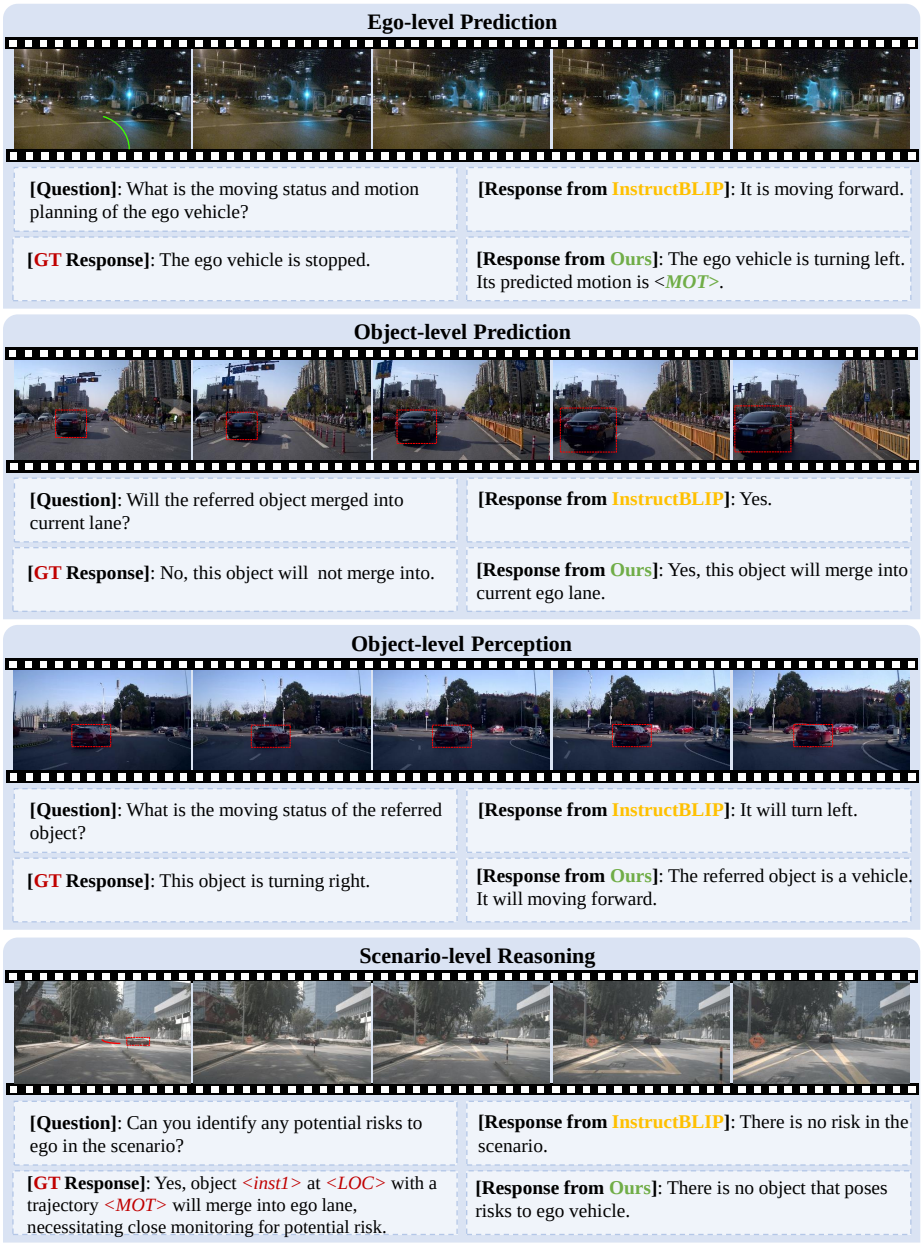

Fig. 4: Failure examples. Locations and motions are pictured in the first frame for better visualization. Ground truth in red color and prediction in green color.