Automated Evaluation of Large Vision-Language Models on Self-driving Corner Cases

Kai Chen[1*] Yanze Li[2*] Wenhua Zhang[2*] Yanxin Liu[2] Pengxiang Li[2] Ruiyuan Gao[3]
Lanqing Hong[4†] Meng Tian[4] Xinhai Zhao[4] Zhenguo Li[4]
Dit-Yan Yeung[1] Huchuan Lu[2] Xu Jia[2†]
[1]Hong Kong University of Science and Technology [2]Dalian University of Technology
[3]The Chinese University of Hong Kong [4]Huawei Noah's Ark Lab
https://coda-dataset.github.io/coda-lm/

Abstract

*Large Vision-Language Models (LVLMs) have received widespread attentions for advancing the interpretable self-driving. Existing evaluations of LVLMs primarily focus on multi-faceted capabilities in natural circumstances, lacking automated and quantifiable assessment for self-driving, let alone the severe road corner cases. In this work, we propose **CODA-LM**, the very first benchmark for the automatic evaluation of LVLMs for self-driving corner cases. We adopt a hierarchical data structure and prompt powerful LVLMs to analyze complex driving scenes and generate high-quality pre-annotations for the human annotators, while for LVLM evaluation, we show that using the text-only large language models (LLMs) as judges reveals even better alignment with human preferences than the LVLM judges. Moreover, with our CODA-LM, we build **CODA-VLM**, a new driving LVLM surpassing all open-sourced counterparts on CODA-LM. Our CODA-VLM performs comparably with GPT-4V, even surpassing GPT-4V by **+21.42%** on the regional perception task. We hope CODA-LM can become the catalyst to promote interpretable self-driving empowered by LVLMs.*

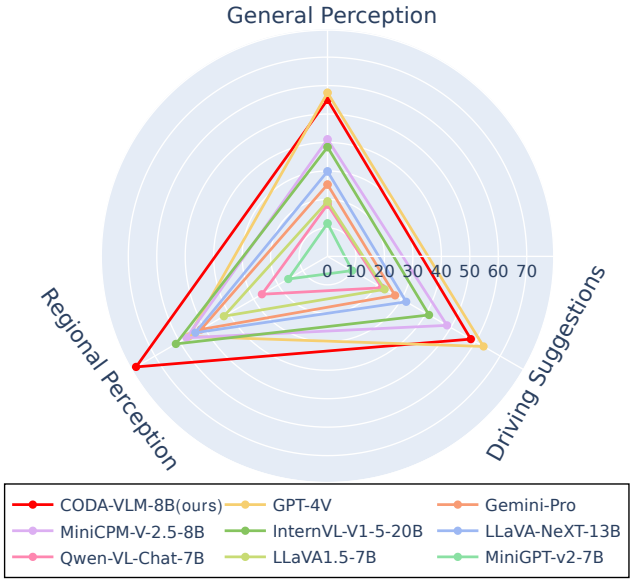

Figure 1. **Comparison among open-sourced and commercial LVLMs on CODA-LM**. CODA-LM provides the very first automated and quantifiable evaluation of LVLMs on road corner cases.

1. Introduction

Large Vision-Language Models (LVLMs) [8, 19, 32, 39] have attracted increasing attention, primarily due to their remarkable visual reasoning abilities, which are of paramount importance [23, 42] for the autonomous driving. Traditional self-driving systems use a modular design, integrating various modules including perception, prediction, and planning to handle complicated road scenarios, which, however, are still inadequate to generalize in the open domain, especially for the severe real-world *corner cases* [28]. In this paper, we primarily consider *object-level corner cases*[1], including both the *instances of novel categories* and *novel instances of common categories* [28].

LVLMs, on the other hand, with their extensive world knowledge and reasoning capability, have the potential to overcome these severe challenges. A preliminary study [48] has revealed the capability of powerful LVLMs [39] in handling the road corner cases, where the samples are selected from CODA [28], the largest real-world corner case dataset, to prompt GPT-4V. Although effective, their evaluation relies on redundant manual inspections, hindering the scalability of larger-scale LVLM evaluation for self-driving.

* Equal contribution. † Corresponding authors.
Contact: jiayushenyang@gmail.com

[1]We adopt the definition of object-level corner case in [6].

Dataset	Multimodal	Corner	General Perception	Regional Perception	Driving Suggestion
CODA [28]	✗	✓	✓	✓	✗
StreetHazards [21]	✗	✓	✓	✓	✗
nuScenes-QA [41]	✓	✗	✓	✗	✗
BDD-X [25]	✓	✗	✓	✗	✗
DRAMA [37]	✓	✗	✓	✓	✓
DriveLM [42]	✓	✗	✓	✓	✓
CODA-LM (ours)	✓	✓	✓	✓	✓

Table 1. **Comparison between CODA-LM and existing datasets**. CODA-LM is the first large-scale multimodal road corner case dataset for interpretable autonomous driving with an automatic and hierarchical evaluation framework.

In this paper, we propose the **CODA-LM**, the very first benchmark for the automated and systematic evaluation of LVLMs on the self-driving corner cases. Following Wen *et al*. [48], we utilize the corner cases from CODA and collect question-answering annotations of three distinct tasks including *general perception*, *regional perception*, and *driving suggestions*. To obtain high-quality pre-annotation, we design a hierarchy data structure to help GPT-4V better analyze complex road scenes and capture all necessary obstacles. The structured responses are then converted to coherent texts, which are then verified by human annotators. Different from the existing LVLM benchmarks [27], we show the necessity of using the text-only LLMs [38] as "judges" for automated evaluation of LVLMs on CODA-LM, which reveals a stronger consistency with humans than LVLM judges [39]. Moreover, we propose **CODA-VLM**, a novel driving LVLM achieving the state-of-the-art among all open-sourced LVLMs on CODA-LM, even surpassing GPT-4V on the regional perception task by **+21.42%**.

The main contributions of this work contain three parts:

1. We propose **CODA-LM**, the **very first** LVLM benchmark for the automatic and systematic evaluation of LVLMs on road corner cases.

2. We demonstrate that text-only LLMs can serve as powerful judges to evaluate LVLMs, revealing a stronger consistency with the human judgments even compared with LVLM judges.

3. We comprehensively assess the performance of existing LVLMs on self-driving corner cases, and construct **CODA-VLM**, a new driving LVLM comparable with GPT-4V on CODA-LM, surpassing all open-sourced counterparts on both perception and suggestions.

2. Related Work

LVLM evaluation primarily focuses on natural image spaces. MME [15] introduces manually designed question-answering pairs to measure both perception and cognition capabilities on a total of 14 sub-tasks. MMBench [33] employs GPT-4 to transform free-form predictions into predefined multiple-choice questions and introduces the CircularEval strategy for a more robust evaluation. The SEED-Bench-2 [26] adopts a similar format with MMBench but extends over 27 dimensions, evaluating LVLMs' abilities in image and text comprehension and interleaved image-text understanding and generation tasks. Auto-Bench [24] generates question-answer-reasoning triplets using LLMs [11, 18, 35, 45] as evaluation data. Tri-HE [49], instead, focuses on LVLM hallucination with a unified evaluation framework. All the evaluation benchmarks above rely on the rigid, manually curated datasets of natural images, and thus, difficult to apply for complicated driving scenarios.

Autonomous driving datasets. The NuScenes-QA [41] manually constructs 460K question-answer pairs based on the object attributes and relationships among objects in scene graphs. BDD-X [25] focuses on the behavior of the ego car and provides corresponding reasons. While both datasets concentrate on general perception, DRAMA [37] and DriveLM [42] further consider regional perception and driving suggestions. DRAMA identifies the most critical targets and offers the corresponding advice, while DriveLM promotes end-to-end autonomous driving understanding through the usage of graph-structured question-answer pairs. Self-driving systems often fail in corner cases, leading to severe accidents. StreetHazards [21] is a synthesized dataset where corner cases are simulated via graphics. CODA [28] is a real-world road corner case dataset with 10K driving scenes, spanning more than 40 classes. As in Tab. 1, the existing corner case datasets lack language modality, while vision-language datasets don't cover road corner cases. Thus, we propose CODA-LM, the first large-scale multimodal road corner case dataset for self-driving with a hierarchical automatic evaluation framework.

3. CODA-LM Dataset

Based on the road corner cases from CODA [28], our CODA-LM comprises 9,768 real-world driving scenarios

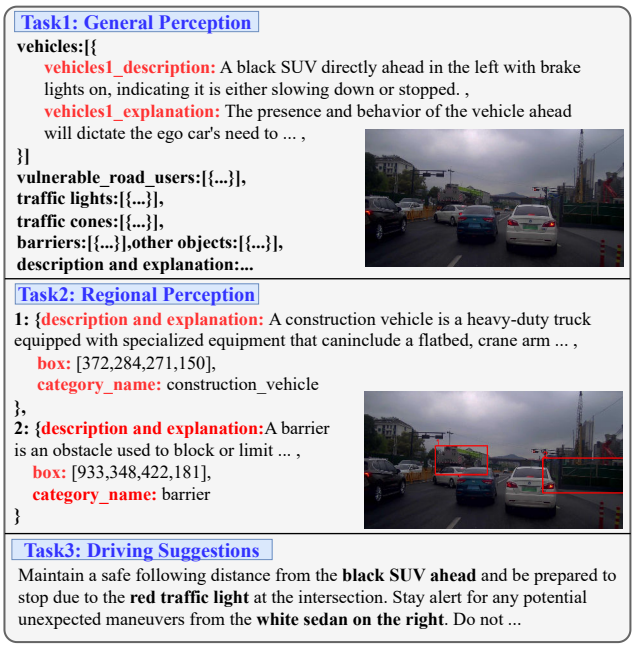

Figure 2. **Task hierarchy of our CODA-LM**, including *general perception* (up), *regional perception* (middle), and *driving suggestions* (bottom), respectively.

with 41,722 textual annotations for critical road entities and 21,537 annotations for road corner cases. Critical road entities affecting self-driving decision-making are categorized into seven distinct groups, including *vehicles*, *vulnerable road users (VRUs)*, *traffic signs*, *traffic lights*, *traffic cones*, *barriers*, and *other objects* (*e.g.*, animals and traffic islands). As illustrated in Fig. 2, our CODA-LM involves a task hierarchy with three principal tasks, including the *general perception*, *regional perception*, and *driving suggestion*, as detailed in Sec. 3.1-3.3 separately. Such a systematic task hierarchy requires LVLMs to understand complex driving situations, providing a comprehensive assessment of **interpretable** self-driving agents empowered by LVLMs.

3.1. General Perception

The foundational aspect of the general perception task lies in a comprehensive understanding of critical road key entities in driving scenarios, including their appearance, location, and reasons why they influence the driving behaviors of our ego car. This task is pivotal in evaluating LVLMs' proficiency in interpreting complex interactive scenes, mirroring the perception process in self-driving. Moreover, to comprehensively evaluate LVLMs' performance in different environments, we classify the images based on the time and weather conditions, including *night* and *daytime* scenes for time conditions, as well as *clear*, *cloudy*, and *rainy* circumstances for the weather conditions.

3.2. Regional Perception

The regional perception task measures LVLMs' capabilities to understand corner case objects when provided with specific bounding boxes, which involves describing objects within the given bounding boxes and explaining why they would influence self-driving behavior. The establishment of regional perception is based on a core realization [20] that the ability to accurately localize corner cases is crucial for enhancing the overall system's robustness in the practical application of autonomous driving. These scenarios often contain complicated or unusual elements that traditional models might overlook or struggle to interpret correctly, such as *unique traffic signs*, *pedestrians with abnormal behavior*, and *atypical road conditions*. By specifically focusing on these cases, we can gain a comprehensive understanding of LVLMs' ability to comprehend corner cases.

3.3. Driving Suggestions

The driving suggestions task aims to evaluate the capability of LVLMs in formulating driving advice, a critical component for interpretable self-driving. This task is closely related to the planning process of autonomous driving, requiring the model to provide the optimal driving suggestions for the ego car after correctly perceiving the general and regional aspects of the current driving environment. Via the construction of the driving suggestions task, we can deeply evaluate the performance of LVLMs in formulating effective driving strategies.

4. CODA-LM Construction

4.1. Data Collection

Overview. For each task introduced in Sec. 3, we meticulously design prompts to guide GPT-4V [2] to generate high-quality textual pre-annotations based on visual information, as provided in Figs. 5 and 6. We start by constructing a hierarchical data structure in the JSON format (detailed in the following) to guide GPT-4V for better scene understanding of complex road scenes, categorizing the critical road entities into seven classes. Each entity is detailedly described, explaining how they affect the driving behavior of the ego car. After obtaining the GPT-4V responses for both the general and regional perceptions, we combine these with the corresponding road image to form a composite context for the GPT-4V to generate the driving suggestions. Finally, we ask human annotators to verify and revise the pre-annotations. The construction pipeline is shown in Fig. 3.

Hierarchical text structure for general perception. To conduct precise perception and even driving suggestions, it is essential to recognize all road obstacles. However, if

[2]https://chatgpt.ust.hk

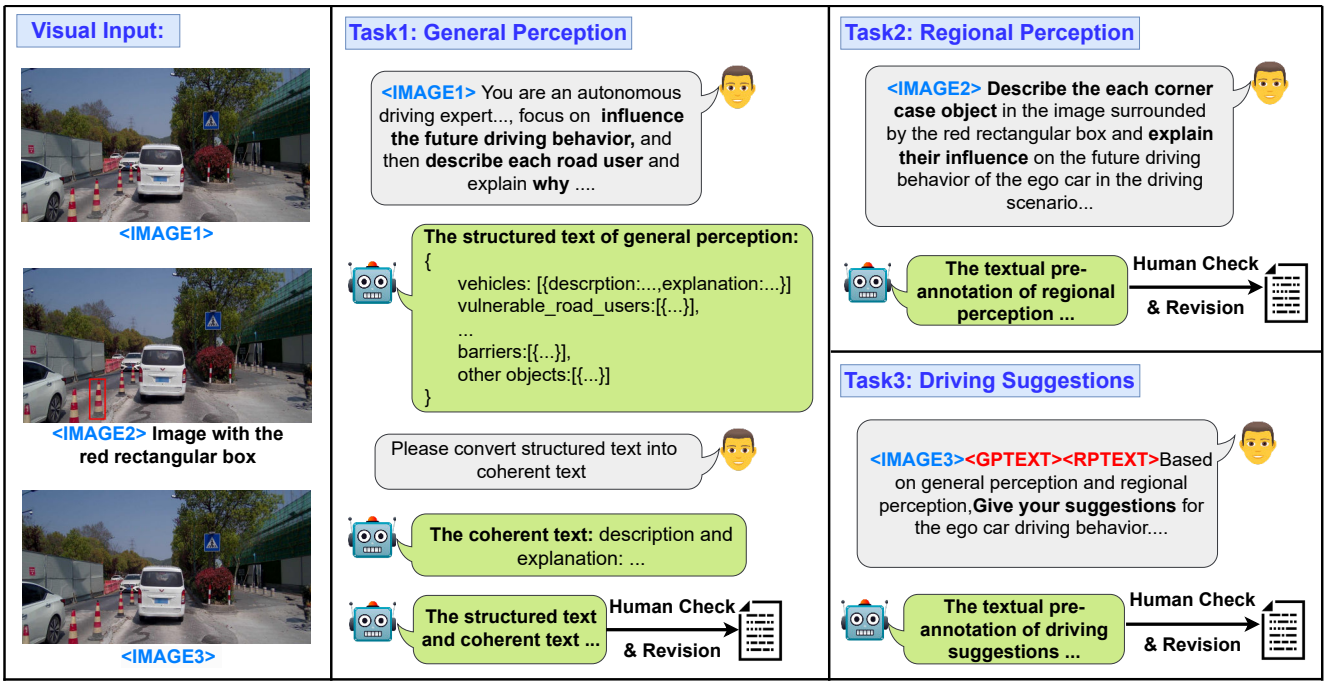

Figure 3. **Overview of CODA-LM construction.** We design a hierarchical data structure in the JSON format to guide GPT-4V to better understand complicated driving scenes and generate high-quality pre-annotations for human annotators to conduct further verification and revision. <GPTEXT> and <RPTEXT> refer to the revised answer from the general perception and the regional perception, respectively.

directly prompted with plain texts, we notice that GPT-4V suffers from 1) *entity ignorance*: GPT-4V tends to focus on the salient objects while ignoring the insignificant obstacles. 2) *element ignorance*: when prompted with plain texts, GPT-4V might describe road entities without explaining why it affects the ego car or vice versa.

Thus, as in Fig. 3 (middle), we design a hierarchy data structure in the JSON format from *categories* to *objects* and ultimately *data elements*. GPT-4V is guided to first recognize objects of every single *category* separately, and "fill in" *description* and *explanation* of each object. We then prompt GPT-4V again to convert structured texts to coherent natural languages and serve as the final pre-annotations. As in Tab. 4, the **"structure-coherence"** pipeline achieves significant consistency with humans.

Visual prompts for regional perception. We consider two manners to convert bounding boxes as the inputs for LVLMs, 1) *visualization*: suggests marking the targets with red rectangle boxes on the original images, as in Fig. 3 (left). 2) *grounding*: uses normalized coordinates (top-left and bottom-right corners) in text prompts to locate the target, similarly with LLaVA [32]. As in Tab. 5, visualization with red rectangles reveals significantly better empirical results, which is considered as the default vision prompts.

Human verification and revision is ultimately adopted to guarantee the correctness of our CODA-LM annotations. For convenience, we construct a labeling tool GUI based on Gradio [1], as in Fig. 9, followed by the ethics review.

Data split. We separate 4,884 scenes as the training set, with 4,384 data samples as the validation set and the remaining 500 samples as the test set to construct the CODA-LM benchmark as in Tab. 2 for a comprehensive comparison among LVLMs.

4.2. Evaluation Framework

Unsatisfactory LVLM judges. LMSYS [50] shows the feasibility of using GPT-4 as judges to evaluate the intelligent chat assistants by giving a 1-10 score, revealing high consistency with human assessment. Inspired by that, we start with a preliminary attempt by using *LVLM judges* (*e.g.*, GPT-4V) to evaluate various LVLMs, which, however, merely obtains a human consistency of around 70% for all three tasks, as shown in Tab. 3.

We assume that this is probably due to the unsatisfactory instruction-following ability of GPT-4V, which cannot always respond in the required format [4]. Meanwhile, GPT-4V still lacks the multimodal in-context learning ability, making few-shot evaluation indispensable in complex and varied autonomous driving scenarios.

Text-only LLM as LVLM judges. In this paper, we propose to adopt *text-only LLMs* (*e.g.*, GPT-4) as judges to

evaluate LVLMs on driving scenarios. Given the reference ground truths and few-shot ICL examples, GPT-4 is instructed to evaluate the correctness of model responses with a score ranging from 1 to 10. The average score of the whole evaluation set serves as the final `Text-Score`. We provide the evaluation prompts and ICL examples in Figs. 7 to 11. As shown in Tab. 3, the text-only GPT-4 judge evaluates more consistently with human judgments than the GPT-4V judge.

Potential bias and hallucination To revise that, we ask the human annotators to verify and revise the evaluation results given by GPT-4 and finally report results in Tab. 2.

Evaluation criteria of the general perception include *accuracy*, *hallucination penalty*, and *consistency*. Accuracy evaluates how well LVLMs match with reference ground truths, while the hallucination penalty suggests that LVLMs should not mention entities not collected in the reference, which, otherwise, should be penalized when computing scores. Consistency focuses on the relationship between the object description and the explanation of why it affects the ego car. For driving suggestions, the criteria focus on the *rationality*, *relevance*, and *detail level* of driving suggestions generated by LVLMs. Especially for driving suggestions, we require responses to be specific and actionable, rather than vague or overly broad. Prompts are listed in Fig 7.

Evaluation metrics. As previously introduced, we utilize the `Text-Score` [50] given by text-only GPT-4 judge as the primary evaluation metrics for all three tasks. We further explore the usage of traditional text-generation evaluation metrics as in Tab. 7, which, however, cannot well differentiate the capabilities of various LVLMs under complicated self-driving scenarios.

4.3. CODA-VLM

In this section, we explore improving the performance of LVLM models on road corner cases from the perspectives of both the visual representation and knowledge transfer and construct our **CODA-VLM**, a novel driving LVLM achieving state-of-the-art recognition and planning performance on autonomous driving scenarios.

Knowledge transfer. To acquire more comprehensive pre-training knowledge, we use the LLaVA-Llama-3-8B-v1.1 developed by Xtuner[3] as our baseline, which follows the basic architecture of LLaVA1.5 [30], while replacing the LLM with LLaMA3-8B[4], and performing modality alignment and instruction fine-tuning on a larger dataset. Based on that, we inject knowledge specific to driving scenarios via instruction fine-tuning. Specifically, we organize the image-text pairs from CODA-LM into a dialogue format and employ a rational data sampling strategy to form an instruction-following dataset. Furthermore, to efficiently learn while preserving as much pre-training knowledge as possible, we use LoRA [22] to fine-tune both the LLM and the visual encoder[5].

Visual representation. To obtain more effective visual representations and enhance the model's regional perception capabilities, we refer to the dynamic high resolution (*i.e.*, AnyRes) from LLaVA-NeXT [31]. While retaining the fixed global image resolution, we split original images into different sub-images, each independently encoded by a shared visual encoder, and finally concatenate all visual tokens together before feeding into LLMs. Moreover, considering context lengths and training costs of LLMs, we observe that a 2×2 MaxPool operation on visual tokens of sub-images can effectively reduce redundancy, achieving a better trade-off between efficiency and performance.

Implementation details. It is worth noting that our approach is simple yet effective. The training of CODA-VLM requires only 3 hours on 8 A800 GPUs. Specifically, we use LoRA with r = 256 and α = 256 for the LLM, and r = 64 and α = 16 for the visual encoder, fine-tuning with a context length of 4096. The learning rate is set to $2e^{-4}$, training for 4 epochs with a batch size of 16 per GPU. We utilize the combination of the train and validation splits of CODA-LM, as discussed in Sec. 4.1. In Sec. 5.2 and 5.3.4, we provide more detailed analysis and empirical ablation results on CODA-VLM.

5. CODA-LM Benchmark

In this section, based on the proposed CODA-LM dataset, we start by comparing and analyzing the performance of different LVLMs in Sec. 5.1, followed by an in-depth analysis of model architecture designs in Sec. 5.2. We then conduct an ablation study on critical components of dataset construction and evaluation in Sec. 5.3.

5.1. Main Results

Baselines. In this work, we evaluate a total of 10 LVLMs, including both open-sourced and commercial models. Commercial models consist of the Gemini-Pro [43] and GPT-4V [39], while the open-sourced LVLMs are categorized based on the parameter sizes of their language models. The 7B/8B variants include the MiniGPT-v2 [7], Shikra [13], LLaVA1.5 [30], Qwen-VL-Chat [3] and MiniCPM-Llama3-V-2.5 [44], while the 13B/20B

[3]https://github.com/InternLM/xtuner
[4]https://huggingface.co/meta-llama/Meta-Llama-3-8B
[5]https://huggingface.co/openai/clip-vit-large-patch14-336

Method	General↑	Regional Perception ↑								Suggestion↑
	Text-Score	ALL	Vehicle	VRU	Sign	Light	Cone	Barrier	Other	Text-Score
MiniGPT-v2-7B	11.58	15.93	18.74	13.58	15.71	17.78	15.34	13.02	14.41	10.00
Shikra-7B	12.24	22.94	28.29	17.88	20.00	15.56	21.23	20.00	19.67	10.20
LLaVA1.5-7B	19.30	42.06	46.67	38.47	39.14	48.89	50.83	30.93	33.82	23.16
Qwen-VL-Chat-7B	18.22	26.62	35.48	24.16	20.86	23.33	19.61	17.56	25.86	22.06
MiniCPM-V-2.5-8B	41.12	57.20	61.91	54.82	59.43	46.67	66.57	35.35	58.75	48.48
LLaVA1.5-13B	24.54	42.41	53.62	36.79	33.71	46.67	41.27	30.41	33.82	27.90
LLaVA-NeXT-13B	29.86	53.63	55.51	47.08	54.00	60.00	70.34	40.47	46.45	31.92
InternVL-V1-5-20B	38.38	61.53	63.77	53.14	50.57	57.78	80.34	46.86	57.11	41.18
Gemini-Pro	25.24	51.38	49.03	42.77	37.43	42.22	69.56	45.70	51.32	27.40
GPT-4V	**57.50**	56.26	60.89	40.58	49.43	54.44	66.08	50.17	53.16	**63.30**
CODA-VLM (ours)	55.04	**77.68**	**78.79**	**73.80**	**64.86**	**73.33**	**86.18**	**78.72**	**68.75**	58.14

Table 2. **Comparison among open-sourced and commercial LVLMs on CODA-LM Test set.** All open-sourced LVLMs suffer from the complicated road corner cases, while our **CODA-VLM**, due to its usage of superior vision representation and knowledge transfer, performs the best or second best on all evaluated dimensions, surpassing all open-sourced counterparts. Note that here we re-scale the original 1-10 `Text-Score` to 1-100 for better readability. **Bold** denotes the best results, while underline suggests the second best.

LVLMs consist of LLaVA1.5 [30], LLaVA-NeXT [31] and InternVL-Chat-V1-5 [14]. Each model is evaluated on the three tasks separately for a comprehensive analysis of their performance on self-driving corner cases.

Setting. To ensure the reproducibility of our evaluation results, we use the same prompt for generating responses for all evaluated LVLMs and employ greedy decoding during inference, which generates the next token with the highest probability at each step as output, thus eliminating randomness during inference. As discussed in Sec. 4.2, GPT-4 is used as the judge for evaluation, with the temperature coefficient set to 0 and a fixed random seed, to ensure consistency when scoring different models.

Results. The comparison results on the CODA-LM Test set are reported in Tab. 2. Among the open-sourced baselines, MiniCPM-V-2.5-8B achieves the best performance, probably due to the usage of the powerful LLaMA3 base model, only ranking second to Intern-VL-1.5-20B on regional perception. Among the commercial models, GPT-4V continues to demonstrate a leadership position, ranking first on general perception and driving suggestions. Interestingly, Gemini-Pro is polarized, showing poor results in general perception and driving suggestions while excelling in regional perception. **CODA-VLM**, instead, achieves the best or second best on all the evaluated dimensions, surpassing all open-sourced counterparts. CODA-VLM obtains comparable performance with GPT-4V, even exceeding GPT-4V by **+21.42%** on regional perception. A qualitative comparison is given in Fig. 4.

5.2. Analysis

Visual represention. Recent works [14,31] have revealed the significant benefit of utilizing high-resolution images as input for LVLMs. For regional perception, simply increasing the image resolution from 224 to 336 enables LLaVA1.5-7B to outperform Shikra-7B by 20%. By further increasing the effective resolution with the AnyRes, LLaVA-NeXT-13B surpasses the LLaVA1.5-13B by over 11%. The compression of visual tokens is another factor. Even with a 448 image resolution, Qwen-VL-Chat-7B is 16% lower than LLaVA1.5-7B with 336 image inputs, largely due to the usage of Q-former for token compression. In contrast, InternVL-V1-5-20B merges four adjacent tokens, while MiniCPM-LLaMA3-V-2.5 resamples each sub-image individually, both effectively reducing redundant tokens while maximizing performance retention. The same tendency can be observed in general perception and driving suggestions tasks. Therefore, in CODA-VLM, we adopt AnyRes with a 2×2 MaxPool to achieve the balance between performance and efficiency.

Knowledge transfer. The knowledge embedded in LVLMs significantly influences the performance, which, on the one hand, comes from pre-trained visual encoders and LLMs, while on the other hand, also arises from high-quality visual instruction fine-tuning. As reported in Tab. 2, MiniCPM-V-2.5-8B surpasses LLaVA-NeXT-13B by 12% and 17% in general perception and driving suggestions, despite having smaller LLMs, revealing the significance of LLaMA3-8B. Moreover, we observe that GPT-4V exceeds open-sourced LVLMs by a significant margin on general

Judge	Reference	General	Regional	Suggestion
GPT-4	GT	**83.67**	**85.71**	**89.80**
GPT-4V	Image	69.39	75.51	69.39
GPT-4V	Img & GT	79.59	79.59	87.76

Table 3. **Consistency between different judges and human judgments.** Text-only GPT-4 judges reveal superior consistency for all tasks. GT denotes ground truth answers. Default settings are marked in gray.

Judge	Reference	Consistency (%)
GPT-4	Plain	71.43
GPT-4	Structured & Concat	77.55
GPT-4	Structured & Coherent	**83.67**

Table 4. **Consistency among human judgments and GPT-4 judges with different references.** The *structured coherence* manner reveals significant superiority.

Method	Grounding	Visualization
Shikra-7B	20.39	**22.94**$^{+2.55}$
LLaVA1.5-13B	18.41	**42.41**$^{+24.0}$
GPT-4V	12.85	**56.26**$^{+43.41}$

Table 5. **Ablation on visual prompts for regional perception.** Visualization with red rectangle boxes shows consistent improvements among all evaluated models.

Model	Training Time	General Perception	Driving Suggestion
LLaVA-1.5	-	15.84	29.24
+ Drive SFT Data	1.5h	53.35	60.83
+ CLIP LoRA	1.6h	53.65	61.17
+ AnyRes	6h	**57.46**	**61.83**
+ 2×2 MaxPool	3h	56.04	61.42

Table 6. **Ablation on our CODA-VLM components.** Training time (hours) is estimated with 8× A800 GPUs.

perception and driving suggestions, indicating that current open-sourced LVLMs still lack the domain-specific knowledge of self-driving. Therefore, in CODA-VLM, we adopt LLaMA3-8B as our base model and conduct the domain-specific fine-tuning with driving scenes in CODA-LM.

5.3. Ablation Study

5.3.1 Human Consistency of Judges

Following LMSYS [50], we adopt the ranking-based manner to calculate the consistency of the GPT-4 and GPT-4V judges with human judgments. We randomly sample 50 samples from the CODA-LM Test set, and for each sample, we further sample two model responses from Tab. 2, followed by random shuffling. We then ask judges to determine the ranking (with ties) of the two candidate responses and human consistency is calculated as the probability of the GPT judge agreeing on the ranking with human judgments.

As reported in Tab. 3, the text-only GPT-4 judge with the reference answers achieves more than 80% consistency for all three tasks, surpassing the GPT-4V variants by a large margin. The GPT-4V judge suffers when only images are provided as the reference, which is relieved when reference answers are provided, but still inferior to the text-only GPT-4 judge, even with a higher expense.

5.3.2 Hierarchical Data Structure for General Perception

We ablate the necessity of using the "structured-coherence" pipeline in Tab. 4. Following Sec. 5.3.1, we evaluate the quality of pre-annotations by using them as the reference for the GPT-4 judge and then calculate the consistency with human judgments. We compare with 1) plain text prompting and 2) structured prompting followed by concatenating annotations of each category to consecutive texts. As shown in Tab. 4, generating structured responses followed by coherence obtains the best consistency.

5.3.3 Visual Prompts for Regional Perception

We ablate the advantage of using visualization over grounding as visual prompts for regional perception. The prompt for visualization is "*Please describe the object inside the red rectangle in the image and explain why it affects ego car driving*", while the prompt for grounding is "*Please provide a description for this object and explain why this object affects ego car driving: [x1, y1, x2, y2]*". As reported in Tab. 5, visualization demonstrates consistent improvement for all evaluated LVLMs, even for Shikra-7B which has been pre-trained with grounding data specifically.

5.3.4 CODA-VLM Components

We ablate the usage of different components of CODA-VLM on a 200-image subset of the CODA-LM Test set. Starting from a pre-trained LLaMA3-8B-based LLaVA1.5 checkpoint, we ablate the usage of 1) domain-specific fine-tuning, 2) training CLIP encoder with LoRA, 3) adopting AnyRes and 4) conducting 2×2 MaxPool step by step. As shown in Tab. 6, our CODA-VLM achieves a better trade-off among efficiency and performance.

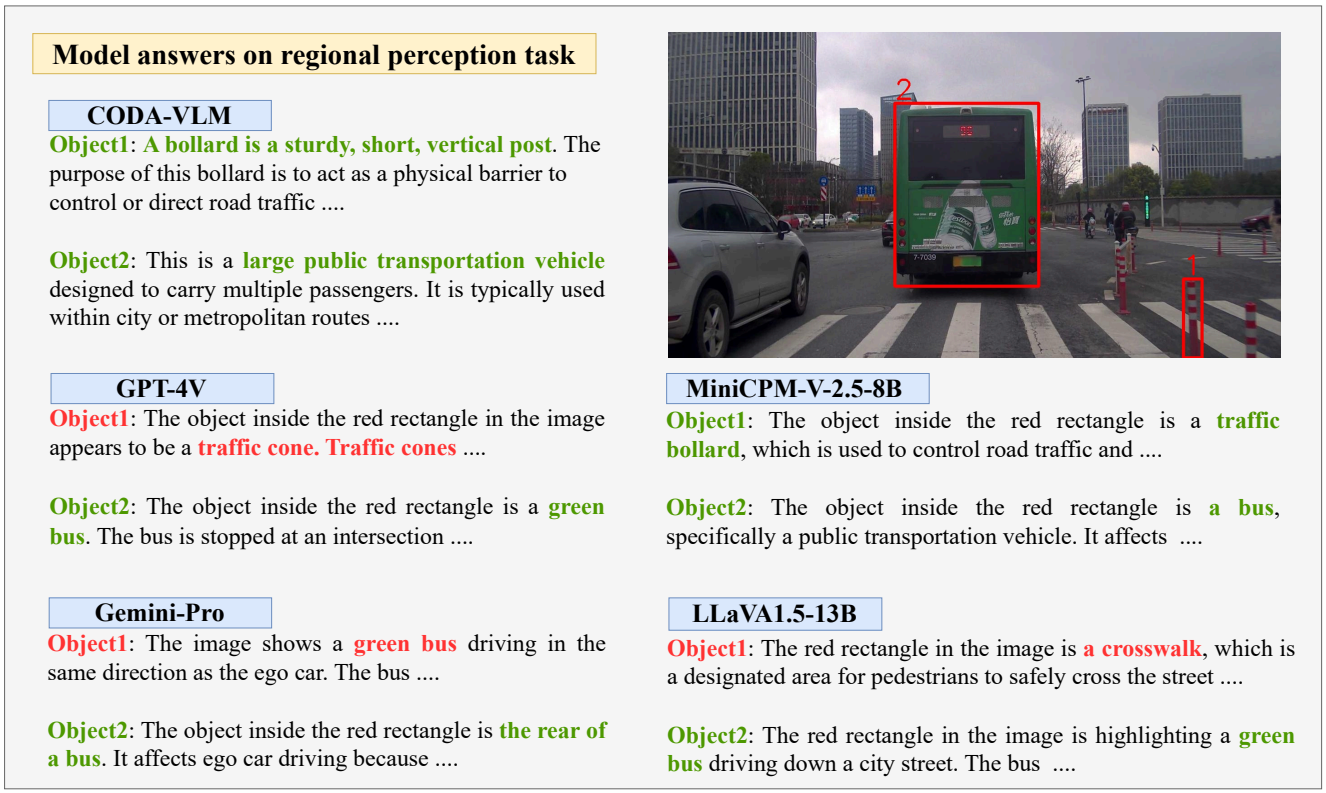

Figure 4. **Qualitative comparison among different LVLMs on the regional perception task.** Mistakes within the model responce are highlighted in **red**, whereas the accurate parts are emphasized in **green**.

5.4. Limitations

CODA-LM is built on corner cases from CODA, which might not cover all possible unexpected conditions in driving scenarios, and we opt to explore controllable generation [12, 16, 17, 29, 34, 47] to generate corner cases in the future. CODA-LM focuses on interpretable self-driving, and we will explore collecting action-level annotations. The current data collection pipeline relies on human verification and revision to ensure the quality of annotations, and an automatic data calibration method is also appealing. How to better incorporate visual pre-trained prior (*e.g.*, self-supervised learning [9, 10, 36, 51]) is also open.

6. Conclusion

In this paper, we propose CODA-LM, a novel real-world multimodality road corner case dataset for autonomous driving with a hierarchy task framework, spanning from general and regional perception to driving suggestions, to support automated evaluation of Large Vision-language Models (LVLMs) on self-driving corner cases. We conduct a comprehensive evaluation of representative LVLMs on road corner cases and propose CODA-VLM, a novel driving LVLM specialized in driving perception and suggestions. However, we are still far from a fully intelligent driving agent and we hope our CODA-LM can serve as the catalyst to promote the development of reliable and interpretable autonomous driving systems.

Acknowledgments. We gratefully acknowledge supports of MindSpore, CANN (Compute Architecture for Neural Networks) and Ascend AI Processor used for this research. The research was partially supported by the National Natural Science Foundation of China (grants Nos. 62472065, U23B2010, and 62106036). This research has been made possible by funding support from the Research Grants Council of Hong Kong through the Research Impact Fund project R6003-21.

References

[1] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*, 2019. 4

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016. 11

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 5

[4] Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv:2308.16890*, 2023. 4

[5] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005. 11

[6] Jasmin Breitenstein, Jan-Aike Termöhlen, Daniel Lipinski, and Tim Fingscheidt. Corner cases for visual perception in automated driving: Some guidance on detection approaches. *arXiv preprint arXiv:2102.05897*, 2021. 1

[7] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 5, 13

[8] Kai Chen, Yunhao Gou, Runhui Huang, Zhili Liu, Daxin Tan, Jing Xu, Chunwei Wang, Yi Zhu, Yihan Zeng, Kuo Yang, et al. Emova: Empowering language models to see, hear and speak with vivid emotions. *arXiv preprint arXiv:2409.18042*, 2024. 1

[9] Kai Chen, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. Multisiam: Self-supervised multi-instance siamese representation learning for autonomous driving. In *ICCV*, 2021. 8

[10] Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. Mixed autoencoder for self-supervised visual representation learning. In *CVPR*, 2023. 8

[11] Kai Chen, Chunwei Wang, Kuo Yang, Jianhua Han, Lanqing Hong, Fei Mi, Hang Xu, Zhengying Liu, Wenyong Huang, Zhenguo Li, et al. Gaining wisdom from setbacks: Aligning large language models via mistake analysis. *arXiv preprint arXiv:2310.10477*, 2023. 2

[12] Kai Chen, Enze Xie, Zhe Chen, Lanqing Hong, Zhenguo Li, and Dit-Yan Yeung. Integrating geometric control into text-to-image diffusion models for high-quality detection data generation via text prompt. *arXiv preprint arXiv:2306.04607*, 2023. 8

[13] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 5, 13

[14] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 6

[15] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 2

[16] Ruiyuan Gao, Kai Chen, Zhihao Li, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive3d: Controllable 3d generation for any-view rendering in street scenes. *arXiv preprint arXiv:2405.14475*, 2024. 8

[17] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. *arXiv preprint arXiv:2310.02601*, 2023. 8

[18] Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation. *arXiv preprint arXiv:2403.09572*, 2024. 2

[19] Yunhao Gou, Zhili Liu, Kai Chen, Lanqing Hong, Hang Xu, Aoxue Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. Mixture of cluster-conditional lora experts for vision-language instruction tuning. *arXiv preprint arXiv:2312.12379*, 2023. 1

[20] Jianhua Han, Xiwen Liang, Hang Xu, Kai Chen, Lanqing Hong, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Xiaodan Liang, and Chunjing Xu. Soda10m: Towards large-scale object detection benchmark for autonomous driving. *arXiv preprint arXiv:2106.11118*, 2021. 3

[21] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. A benchmark for anomaly segmentation. *arXiv preprint arXiv:1911.11132*, 2019. 2

[22] 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 arXiv:2106.09685*, 2021. 5

[23] 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

[24] Yuanfeng Ji, Chongjian Ge, Weikai Kong, Enze Xie, Zhengying Liu, Zhengguo Li, and Ping Luo. Large language models as automated aligners for benchmarking vision-language models. *arXiv preprint arXiv:2311.14580*, 2023. 2

[25] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *ECCV*, 2018. 2

[26] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.17092*, 2023. 2

[27] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, May 2024. 2

[28] Kaican Li, Kai Chen, Haoyu Wang, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, et al. Coda: A real-world road cor-

ner case dataset for object detection in autonomous driving. *arXiv preprint arXiv:2203.07724*, 2022. 1, 2

[29] Pengxiang Li, Zhili Liu, Kai Chen, Lanqing Hong, Yunzhi Zhuge, Dit-Yan Yeung, Huchuan Lu, and Xu Jia. Trackdiffusion: Multi-object tracking data generation via diffusion models. *arXiv preprint arXiv:2312.00651*, 2023. 8

[30] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 5, 6, 13

[31] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. 5, 6

[32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arxiv preprint arxiv:2304.08485*, 2023. 1, 4

[33] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 2

[34] Zhili Liu, Kai Chen, Yifan Zhang, Jianhua Han, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, and James Kwok. Geom-erasing: Geometry-driven removal of implicit concept in diffusion models. *arXiv preprint arXiv:2310.05873*, 2023. 8

[35] Zhili Liu, Yunhao Gou, Kai Chen, Lanqing Hong, Jiahui Gao, Fei Mi, Yu Zhang, Zhenguo Li, Xin Jiang, Qun Liu, et al. Mixture of insightful experts (mote): The synergy of thought chains and expert mixtures in self-alignment. *arXiv preprint arXiv:2405.00557*, 2024. 2

[36] Zhili Liu, Jianhua Han, Kai Chen, Lanqing Hong, Hang Xu, Chunjing Xu, and Zhenguo Li. Task-customized self-supervised pre-training with scalable dynamic routing. In *AAAI*, 2022. 8

[37] Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. Drama: Joint risk localization and captioning in driving. In *WACV*, 2023. 2

[38] OpenAI. ChatGPT, 2023. 2

[39] OpenAI. ChatGPT-4V System Card, 2023. 1, 2, 5, 13

[40] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 11

[41] 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 arXiv:2305.14836*, 2023. 2

[42] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. *arXiv preprint arXiv:2312.14150*, 2023. 1, 2

[43] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 5, 13

[44] MiniCPM-V Team. MiniCPM-V, 2024. 5

[45] 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 arXiv:2302.13971*, 2023. 2

[46] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 11

[47] Yibo Wang, Ruiyuan Gao, Kai Chen, Kaiqiang Zhou, Yingjie Cai, Lanqing Hong, Zhenguo Li, Lihui Jiang, Dit-Yan Yeung, Qiang Xu, and Kai Zhang. Detdiffusion: Synergizing generative and perceptive models for enhanced data generation and perception. *arXiv preprint arXiv:2403.13304*, 2024. 8

[48] Licheng Wen, Xuemeng Yang, Daocheng Fu, Xiaofeng Wang, Pinlong Cai, Xin Li, Tao Ma, Yingxuan Li, Linran Xu, Dengke Shang, et al. On the road with gpt-4v (ision): Early explorations of visual-language model on autonomous driving. *arXiv preprint arXiv:2311.05332*, 2023. 1, 2

[49] Junjie Wu, Tsz Ting Chung, Kai Chen, and Dit-Yan Yeung. Unified triplet-level hallucination evaluation for large vision-language models. *arXiv preprint arXiv:2410.23114*, 2024. 2

[50] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeruIPS*, 2023. 4, 5, 7

[51] LIU Zhili, Kai Chen, Jianhua Han, HONG Lanqing, Hang Xu, Zhenguo Li, and James Kwok. Task-customized masked autoencoder via mixture of cluster-conditional experts. In *ICLR*, 2023. 8