Appendix

A. FutureVQA

A.1. Dataset Creation and Quality control

Our dataset creation aims to provide a benchmark that address VLMs ability in consistant and accurate future reasoning with focus on diverse questions costomized based on different scene. To achieve this we utilize the annotation pipeline operate with both human and AI agent, which to efficiently create the QA.

(1) Human Expert QA Generation and Quality Control: To construct a human-like benchmark dataset with high diversity, we employed five expert annotators to manually generate question-answer pairs based on selected clips from OpenDV-YouTube dataset [62], covering multiple cities with different weathers. Each QA pair was subsequently reviewed and verified by 1–2 annotators to ensure clarity, unambiguity, and answerability based on the given input. Although time-consuming, this process results in a more diverse and naturally phrased QA dataset compared to rule-based or template-driven approaches.

Compared to existing works in the driving domain (see Figure 2), such as nuScenes-QA [47] and DRAMA [42], which rely on rule-based methods, or OmniDrive [55], which uses GPT-generated data to construct large-scale datasets, our benchmark prioritizes diversity and human-like reasoning. While DriveLM-ns [52] incorporates human annotations for prediction and planning tasks, it still follows a rigid and highly structured question format, regardless of the uniqueness of each video clip. As shown in Table 5, despite being smaller in overall size, our dataset provides over 4× more unique questions, nearly 3× larger vocabulary, and over 400× higher type-token ratio (TTR). Notably, more than 95% of our questions appear fewer than 10 times. In contrast, DriveLM contains over 85% of questions repeated more than 10^2 times, over 20% more than 10^3 times, and over 2% more than 10^4 times, without considering the uniqueness of differences in scene content.

(2) AI Quality Control and Multi-option Generation: To minimize typographical errors, we employ GPT-4o to review all QA pairs generated by human annotators and automatically correct any detected typos. Following this, GPT-4o is further used to generate plausible but incorrect answer options based on the ground-truth answers provided by annotators.

To ensure that the resulting multiple-choice questions remain unambiguous—with only one clearly correct answer—each generated QA pair is manually reviewed by human annotators. This final verification step guarantees the quality and clarity of the multi-option format in our dataset.

Dataset	N. Ques.	N. Uniq. Ques.	Vocab. Size	TTR
DriveLM(Pred.)	123k	15	69	4.1×10^{-5}
DriveLM(Pred.&Percep.)	285k	234	150	4.1×10^{-5}
Ours	2.7k	**969**	**433**	$\mathbf{1.8 \times 10^{-2}}$

Table 5. Comparison of question diversity between our dataset and DriveLM. **N. Ques.** denotes the total number of questions; **N. Uniq. Ques.** represents the number of unique questions after de-duplication; **Vocab. Size** is the number of distinct words used in the questions; and **TTR** (Type-Token Ratio) measures lexical diversity, computed as the ratio of unique words to total words. The results highlight its greater linguistic diversity and reduced reliance on fixed templates.

A.2. Evaluation Protocol

To address the limitations of conventional statistical-based metrics, we adopted an option-based answer format for evaluation, where each question has predefined multiple-choice answers (e.g., `A: Yellow`). The models were required to provide the corresponding option label (e.g., `A`) as the answer. See Figure 6 for the prompt and Algorithm 2 for the multi-trials evaluation.

Interestingly, during our experiments, we observed that not all models consistently adhered to this strict answer format. Some models would output answers like `A: Yellow` or simply `Yellow`. To account for this, we relaxed the evaluation criteria to accept both answer formats as correct. However, models like Qwen-VL-7B [5] still struggled to follow the instructions and produced responses such as *"The answer is A"*, *"The answer is A: Yellow"*, or other variations. Since following instructions is an important part of the evaluation, we did not further relax this restriction, which resulted in lower accuracy for these models, as shown in Table 1.

FutureVQA Prompt

Imagine you are looking at the image ***{future_second}*** second after the input frames and answer the following question:
Question: ***{question}***
Options: ***{options}***

Please choose the most appropriate answer from the given options. Respond with the option without any explanation, for example, if the answer is B: Yellow, your answer should be: B

Figure 6. The prompt used to instruct VLMs to predict the future scene and answer the corresponding question.

A.3. VQA Category

To evaluate the diverse reasoning capabilities of VLMs, we classify VQA tasks into the following categories. These cat-

Benchmark	Task	T. Size	Cust. Q	Ans. Type	Mul. Trl.	Mul. C.	Pred.	T-Pred.
nuScenes-QA [47]	Drive VQA	83.3k**	✗	Mixed	✗	✓	✗	✗
BDD-X [33]	Drive Action	2.6k	-	Sentence	✗	✓	✗	✗
DRAMA [42]	Drive VQA	11.6k	✗	Mixed	✗	✗	✗	✗
Rank2Tell [50]	Drive VQA	-	✗	Mixed	✗	✓	✗	✗
OmniDrive [55]	Drive VQA	24k†	✓	Sentence	✗	✓	✓	✗
DriveLM-nS [52]	Drive VQA	73k*	✗	Sentence	✗	✓	✓	✗
MMBench [40]	Gen. I. QA	1.7k	✓	Options	✓	-	-	-
LngVidBench [60]	Gen. V. QA	5.3k	✓	Options	✗	-	-	-
Video-MME [16]	Gen. V. QA	2.7k	✓	Options	✗	-	-	-
MME [15]	Gen. I. QA	2.1k	✗	Y/N	✗	-	-	-
Ours	Drive VQA	2.8k	✓	Options	✓	✓	✓	✓

Table 6. Comparison of existing VLM benchmarks. Key aspects of dataset creation include test size (T. Size), whether questions are customized for different scenarios and video clips (Cust. Q), answer type (Ans. Type), multi-trial evaluation for each question (Mul. Tri), inclusion of multiple cities (Mul. C.), presence of perception tasks (Perc.), inclusion of prediction tasks (Pred.), and whether the dataset challenges VLMs with time-specific prediction (T-Pred.). Our benchmark dataset consists of fully human-annotated QA pairs tailored to different scenes, rather than relying on rule-based methods. Furthermore, our dataset challenges VLMs to predict future scenes at specific time intervals, requiring precise temporal reasoning to differentiate between near-future and far-future events.
†: The QA pairs are fully generated by GPT-4. ** Fully rule-based (no human annotators), * Semi-rule-based labeling (with human annotators for certain tasks).

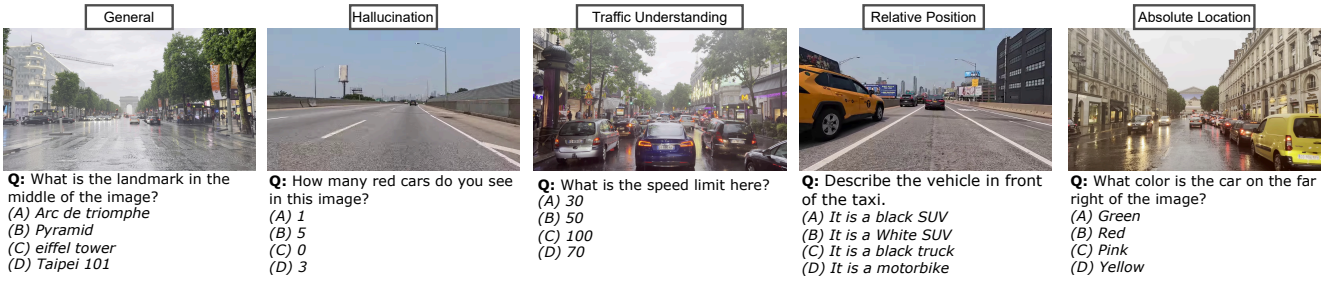
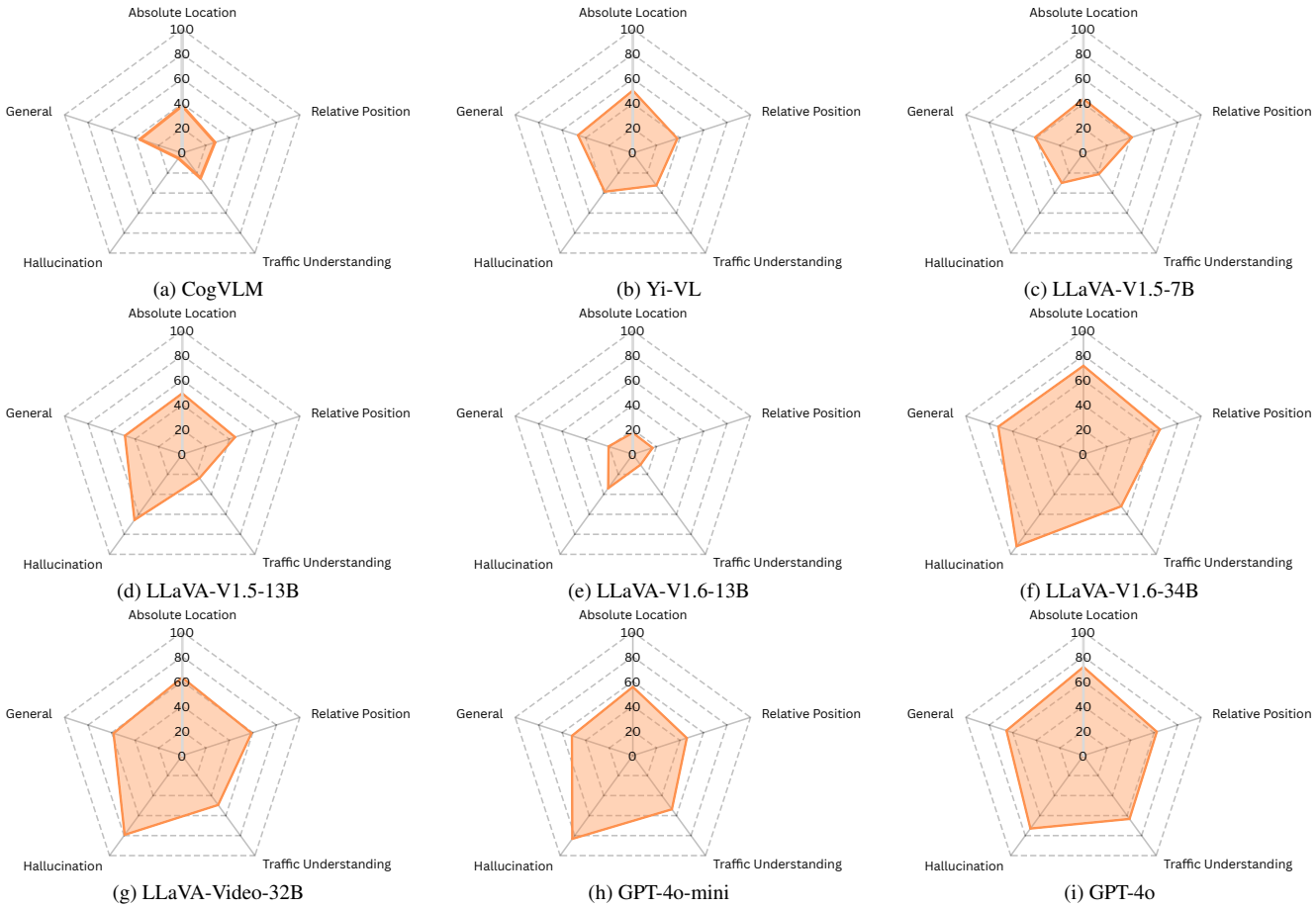


Figure 7. Examples of visual question answering (VQA) tasks categorized into different types: **Hallucination**, **General**, **Traffic Understanding**, **Absolute Location**, and **Relative Position**. Each question is categorized based on the type of reasoning it requires; however, a single question can belong to multiple categories simultaneously, depending on its context and the type of information needed.

egories are not mutually exclusive, as a single question can belong to multiple categories depending on the type of reasoning required.

- **Hallucination:** This category evaluates the model's ability to avoid providing incorrect information about objects or features that do not exist in the scene. (e.g., *"How many blue cars do you see in this image?"*) Such questions are especially challenging when an object has just left the scene.
- **General:** General questions involve straightforward scene understanding or recognition tasks that do not require spatial or temporal reasoning. Examples include identifying landmarks, objects, or common scene elements (e.g., *"What is the landmark in the middle of the image?"*).
- **Traffic Understanding:** This category targets traffic-related reasoning, including understanding road signs, speed limits, or dynamic traffic scenarios. These questions often require knowledge specific to driving environments (e.g., *"What is the speed limit here?"*).
- **Absolute Location:** Absolute location questions focus on the spatial properties of objects in the scene, such as identifying specific positions or attributes relative to the image boundaries (e.g., *"What color is the car on the far right of the image?"*).
- **Relative Position:** Relative position questions require understanding the spatial relationships between multiple objects in the scene. These questions test the model's ability to interpret multiple objects interaction (e.g., *"Describe the vehicle in front of the taxi."*).

By introducing these categories, we aim to provide a comprehensive evaluation framework for VLMs, covering both basic scene understanding and complex reasoning tasks. See Figure 7 for the examples.



(a) CogVLM (b) Yi-VL (c) LLaVA-V1.5-7B

(d) LLaVA-V1.5-13B (e) LLaVA-V1.6-13B (f) LLaVA-V1.6-34B

(g) LLaVA-Video-32B (h) GPT-4o-mini (i) GPT-4o

Figure 8. Radar plots comparing the performance of various models across five VQA categories: Hallucination, General, Traffic Understanding, Absolute Location, and Relative Position. In this experiment, models perform regular VQA on images, with the actual images provided as input. The plots illustrate the strengths and weaknesses of each model in handling different reasoning tasks, providing a comparative baseline for understanding the capabilities of existing VLMs before extending to future image QA tasks.

A.4. Analysis on Different FutureVQA Categories

To establish a baseline for expected performance in the FutureVQA, we analyze various VLMs on our benchmark dataset across different categories . In this baseline analysis, VLMs perform regular VQA, where the actual images corresponding to the questions are provided as input.

As shown in Figure 8, we evaluate models includes CogVLM [56], Yi-VL [65], LLaVA series [37, 39] and GPT-4o, the results suggest that traffic understanding appears to be a relatively weak area for many existing VQA models. Most models do not exhibit significant differences in their capability to handle absolute or relative position questions. Additionally, for hallucination-related tasks, where models are asked about nonexistent objects, most models perform well when the image is provided, effectively avoiding incorrect predictions. These findings highlight the strengths and weaknesses of current VLMs and provide a foundation for evaluating their potential performance in future image QA tasks.

In Figure 9, we further compare the performance of VLMs across different question categories when asked to predict future scenes. As time progresses, we observe that GPT-4o's performance degrades significantly across all categories, with the most notable decline in questions related to relative and absolute positioning.

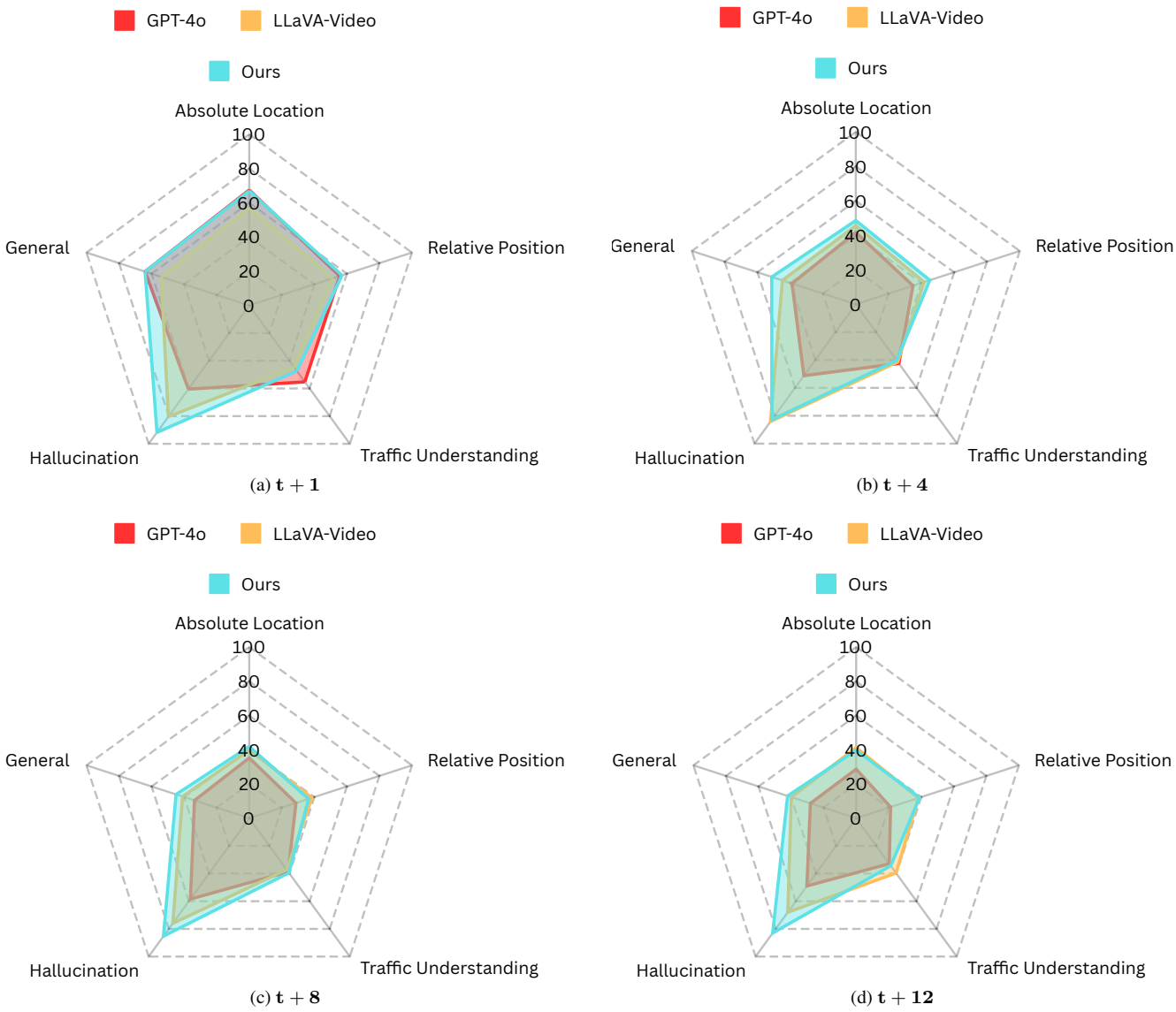


Figure 9. Radar plots comparing the performance of different models across various VQA categories (Hallucination, General, Traffic Understanding, Absolute Location, and Relative Position) at different future time steps: (a) $t + 1$, (b) $t + 4$, (c) $t + 8$, and (d) $t + 12$. The results highlight that while most models maintain robustness in hallucination detection, their performance in other categories, particularly traffic understanding and spatial reasoning, declines as the time offset increases.

B. Detail Implementation

B.1. Chain-of-Thought

To enhance temporal reasoning, we adopt a Chain-of-Thought (CoT) prompting strategy in which the VLM predicts the future scene progressively, one step at a time. Rather than directly predicting the outcome at a future timestamp, the model is encouraged to reason through each intermediate step—first predicting $t = 1$, then $t = 2$, and so on, until the final target time is reached, see Algorithm 3. At each step, the model uses the history frames along with its previous predictions to generate the next future scene description. This design mimics human-like sequential foresight and allows the model to build up an understanding of how the scene may evolve over time. For practical computational efficiency, we limit the maximum number of steps to 4. This step-wise reasoning not only improves temporal consistency but also provides interpretable intermediate predictions that make the model's reasoning process more transparent and grounded in scene dynamics.

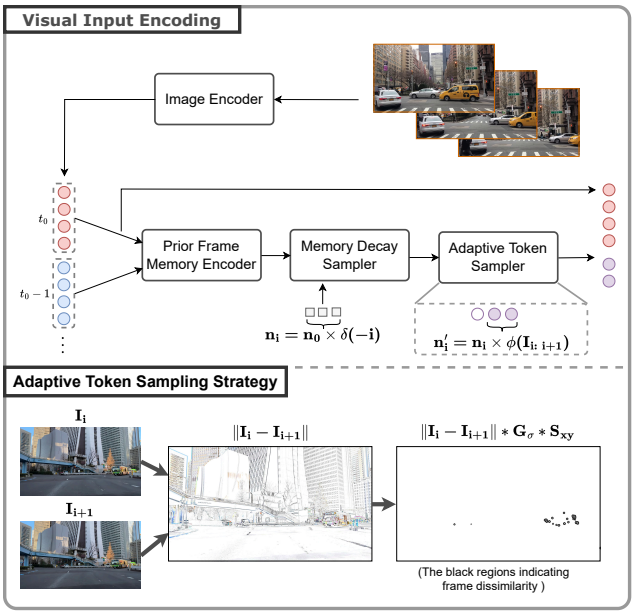


Figure 10. Overview of our visual encoding pipeline. The goal is to minimize the number of tokens while maintaining similar performance. In the context of autonomous driving videos, recent frames typically have greater influence on upcoming events. To reflect this, the **Memory Decay Sampler** assigns fewer queries to older frames, while the **Adaptive Token Sampler** adjusts the number of tokens based on the similarity between adjacent frames. The **Prior Frame Memory Encoder** is a transformer-based module that integrates temporal information from preceding frames.

Figure 11. Visualization of frame similarity evaluation using Gaussian smoothing followed by the Sobel operator with different kernel sizes. The input images have a resolution of 1280 × 720 pixels. The difference between two consecutive frames, $|I(t) - I(t-1)|$, is computed, smoothed using Gaussian filters with kernel sizes of 5, 9, 11, 13, 15, and 17, and then processed with the Sobel operator, S_{xy}, to highlight changes. For better readability, the colors of the similarity maps are inverted.

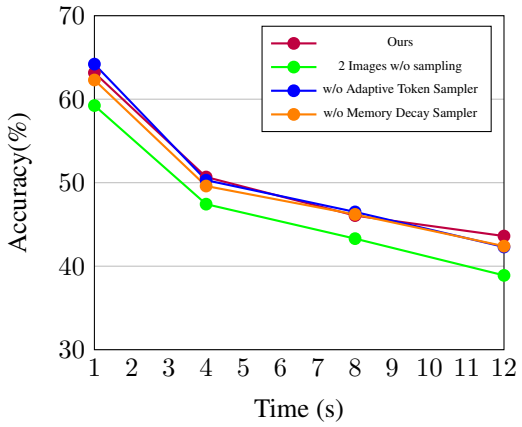


Figure 12. Accuracy over time in the FutureVQA task with different visual input encoding strategies, showing that our sampling approach reduces the number of required tokens while maintaining higher performance.

B.2. Visual Input Encoding

Memory Decay Sampling. Our implementation of the memory decay sampler leverages a transformer-based framework with learnable sampling queries $Q = \{q_1, q_2, \ldots, q_n\}$, where n is the total number of queries set as the initial number of tokens. These queries are initialized at the beginning of training and are optimized to extract temporal information relevant to the task. Let the current time be t_0, and let the number of tokens provided by the image encoder be n_0. The decay factor for the frame at time $t_0 - i$ is defined as $\left(\frac{1}{2}\right)^i$. Accordingly, the first $n_0 \cdot \left(\frac{1}{2}\right)^i$ queries are utilized in the cross-attention mechanism to represent the frame at $t_0 - i$.

Adaptive Token Sampling. In our implementation, frame similarity is evaluated by first computing the difference between two consecutive frames, $|I(t) - I(t-1)|$. To reduce noise introduced by high-frequency details, such as windows on distant buildings in urban environments, a Gaussian filter, G_σ, is applied to smooth the difference map while preserving significant changes. Finally, a Sobel operator, S_{xy}, is used to highlight the structural changes between the frames.

During our experiments, we tested multiple Gaussian filter kernel sizes and determined that a kernel size of 13 strikes the best balance between reducing noise and preserving important structural details. The comparison is shown in Figure 11. After computing the similarity maps, we measure the amount of highlighted area and then scale and cap the values for consistency. On average, the scaling factor is approximately 0.5 across our evaluation dataset.

Time	Model	Scores ↑				
		B-3	B-4	R-L	C	M
+1s	Baseline*	13.4	6.1	25.0	2.4	25.9
	Ours*	**32.2**	**26.2**	**40.2**	**17.7**	**41.5**
	w/o CoT	28.1	22.7	38.2	15.1	40.2
	w/o self-sup.	12.1	7.2	24.9	2.7	26.2
	$t_{0s:-10s}$	32.5	26.5	40.0	17.6	41.3
	w/o Adpt. Sam.	32.3	26.2	40.4	17.3	41.6
	w/o Mem. Sam.	31.8	26.2	40.1	17.0	41.2
+4s	Baseline*	22.5	6.0	24.4	2.6	25.5
	Ours*	**28.6**	**22.5**	**37.1**	**11.8**	**39.1**
	w/o CoT	25.6	20.1	35.9	11.0	38.4
	w/o self-sup.	11.6	6.7	24.4	2.0	25.8
	$t_{0s:-10s}$	32.1	26.2	40.7	17.7	41.5
	w/o Adpt. Sam.	32.7	26.3	40.6	18.0	41.5
	w/o Mem. Sam.	32.7	25.6	40.8	18.1	41.5
+8s	Baseline*	11.2	6.2	25.1	2.2	25.4
	Ours*	**27.5**	**21.4**	**36.2**	**10.1**	**38.3**
	w/o CoT	24.1	18.6	34.9	9.7	37.6
	w/o self-sup.	12.0	7.1	24.9	2.3	26.3
	$t_{0s:-10s}$	32.2	25.8	40.2	17.7	41.5
	w/o Adpt. Sam.	32.3	26.6	41.5	17.5	41.5
	w/o Mem. Sam.	32.0	26.4	41.9	17.1	41.5
+12s	Baseline*	11.3	7.2	23.9	2.2	25.5
	Ours*	**26.7**	**20.6**	**35.6**	**9.4**	**37.7**
	w/o CoT	25.6	20.1	34.4	8.6	36.9
	w/o self-sup.	11.5	6.7	24.4	2.0	25.8
	$t_{0s:-10s}$	31.2	26.0	39.5	17.0	41.5
	w/o Adpt. Sam.	32.0	25.5	39.6	16.9	41.5
	w/o Mem. Sam.	32.1	26.2	40.4	17.9	41.5

Table 7. In this comparison the reference captions are from regular image captioning, while the compared captions are generated by our fine-tuned model which perform future scenes captioning with only previous frames are given. B-3: BLEU-3, B-4: BLEU-4, R-L: ROUGE-L, C: CIDEr, M: METEOR.

C. Additional Evaluation

C.1. Ablation Study on Sampling Strategy

Our choice of the number of visual input frames is guided by two main considerations: (1) performance and (2) hardware constraints. The objective is to minimize the number of visual tokens while maintaining competitive performance. Detailed results at specific time steps are provided in Table 7 and Figure 12.

We observe that extending the input range from the past 5 seconds to the past 10 seconds does not lead to significant performance gains, yet results in increased computational cost. On the other hand, reducing the input to only the past 2 seconds leads to a slight drop in performance.

Similarly, ablating either the memory decay sampler or the adaptive token sampler individually does not substantially affect the final accuracy, while reducing visual token usage by approximately 75%. This highlights the efficiency of our sampling strategy in balancing performance and computational cost.

C.2. Hallucination and Mode Collapse

Generating accurate captions is the first and most crucial step in our training methodology. We experimented with various models for this task; however, we observed that not all models are capable of providing objective and accurate captions that comprehensively describe all elements in the scene. For instance, models like LLaVA-V1.5-7B [38] tend to generate repetitive sentences and often produce hallucinations, exaggerating or inaccurately inferring details that are not present in the image. Figure 13 illustrates examples of these issues, showcasing captions that overstate the number of objects in the scene and use overly similar and redundant phrasing. These limitations highlight the need for more robust captioning models to ensure high-quality data generation for downstream tasks.

D. Prompt Template

In this section, we describe the unified prompt template used for our experiments across three key tasks: captioning evaluation, regular VQA, and FutureVQA. The template, shown in Figure 14, standardizes the model's input format to ensure consistent and fair evaluation.

For captioning evaluation, the model generates captions for a given image, which are subsequently scored by GPT-4o acting as a judge. GPT-4o is instructed to provide a score between 1 and 10 based on the objective aspects of the caption, explicitly disregarding subjective elements such as mood or atmosphere.

In the regular VQA task, the model is provided with the input image and a set of predefined multiple-choice options. It is required to select the most appropriate answer, establishing a baseline for evaluating the model's performance when the image is explicitly available. In Figure 15 and Figure 16, we show that even without explicitly tuning the baseline model, having sufficient knowledge to interpret the visual input does not translate into temporal reasoning ability. The model fails to understand how events unfold over time and cannot align the scene for both near and distant future predictions.

LLaVA-1.5-7B

The image features a busy street with multiple cars and trucks. There are **at least 11 cars and 3 trucks visible on the road**. Some cars are positioned closer to the foreground, while others are further back in the scene. The vehicles are driving under a bridge, and there are traffic lights present in the area. The traffic lights are located at different positions along the road, ensuring proper traffic flow and safety.

The image shows a busy street with multiple cars and trucks. **There are at least 11 cars and 2 trucks visible on the road.** Some cars are positioned closer to the foreground, while others are further back in the scene. The vehicles are driving in various directions, creating a sense of movement and activity on the street.

LLaVA-1.6-34B

The image depicts a highway scene with a focus on a road intersection. There is a white SUV in the foreground, stopped at the intersection. The road is marked with white lane markings and a dashed white line indicating a lane change area. A yellow diamond-shaped sign with a black symbol of a person walking is visible on the right side of the road, indicating a pedestrian crossing or area.

Above the intersection, there is a green highway sign with white lettering that reads ""Left Exit 1A,"" suggesting the direction of the exit ramp. The sign also includes a symbol indicating the exit is to the left.

The highway overpass is in the background, with a greenish-gray color. The sky is clear and blue, indicating fair weather conditions. There are no visible people or animals in the image. The style of the image is a standard, non-stylized photograph with no filters or artistic effects applied. The focus is on the road infrastructure and the vehicle, with no additional context or background details provided.

In the image, there is a highway scene with several vehicles and a clear sky. On the left side of the image, there is a black SUV driving on the highway. The SUV is positioned in the left lane, and it appears to be in motion.

In the center of the image, there is a large black van or bus. It is driving in the right lane of the highway. The vehicle is quite large, occupying a significant portion of the image from left to right.

On the right side of the image, there is a black car driving in the right lane of the highway. The car is smaller in comparison to the van and is positioned behind it.

In the background, there are trees and a clear sky, indicating that the highway is likely in a suburban or rural area. There are no people visible in the image. The overall scene suggests a typical day with vehicles traveling on a highway.

Figure 13. Comparison of captions generated by LLaVA-V1.5-7B [38] and LLaVA-V1.6-34B [39]. While LLaVA-V1.5-7B produces shorter and repetitive captions with occasional hallucination, LLaVA-V1.6-34B generates significantly longer and more detailed descriptions. Additionally, LLaVA-V1.6-34B exhibits a varied response pattern, providing distinct levels of detail and focus when presented with different images.

GPT-4o as Judge for Captioning Evaluation

Please act as an impartial judge and evaluate the quality of the image caption provided by an AI assistant displayed below. Your evaluation should specifically assess the accuracy of object presence and positioning within the image, disregarding any subjective descriptions like vibe, atmosphere, or general impressions. Focus solely on whether the caption correctly reflects the precise positioning and presence of each object mentioned. Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your brief explanation, please rate the response on a scale of 1 to 10 by strictly following this format: '[[rating]]', for example: 'Rating: [[5]]'.
Caption by the AI assistant: ***{caption}***

Regular VQA

Answer the following question based on the image:
Question: ***{question}***
Options: ***{options}***
Please choose the most appropriate answer from the given options. Respond with the option without any explanation, for example, if the answer is B:
Yellow, your answer should be: B

Figure 14. Prompts used for three tasks: GPT-4o as a judge in captioning evaluation and Regular VQA on our annotated evaluation dataset. Each prompt is tailored to the specific requirements of its respective task.

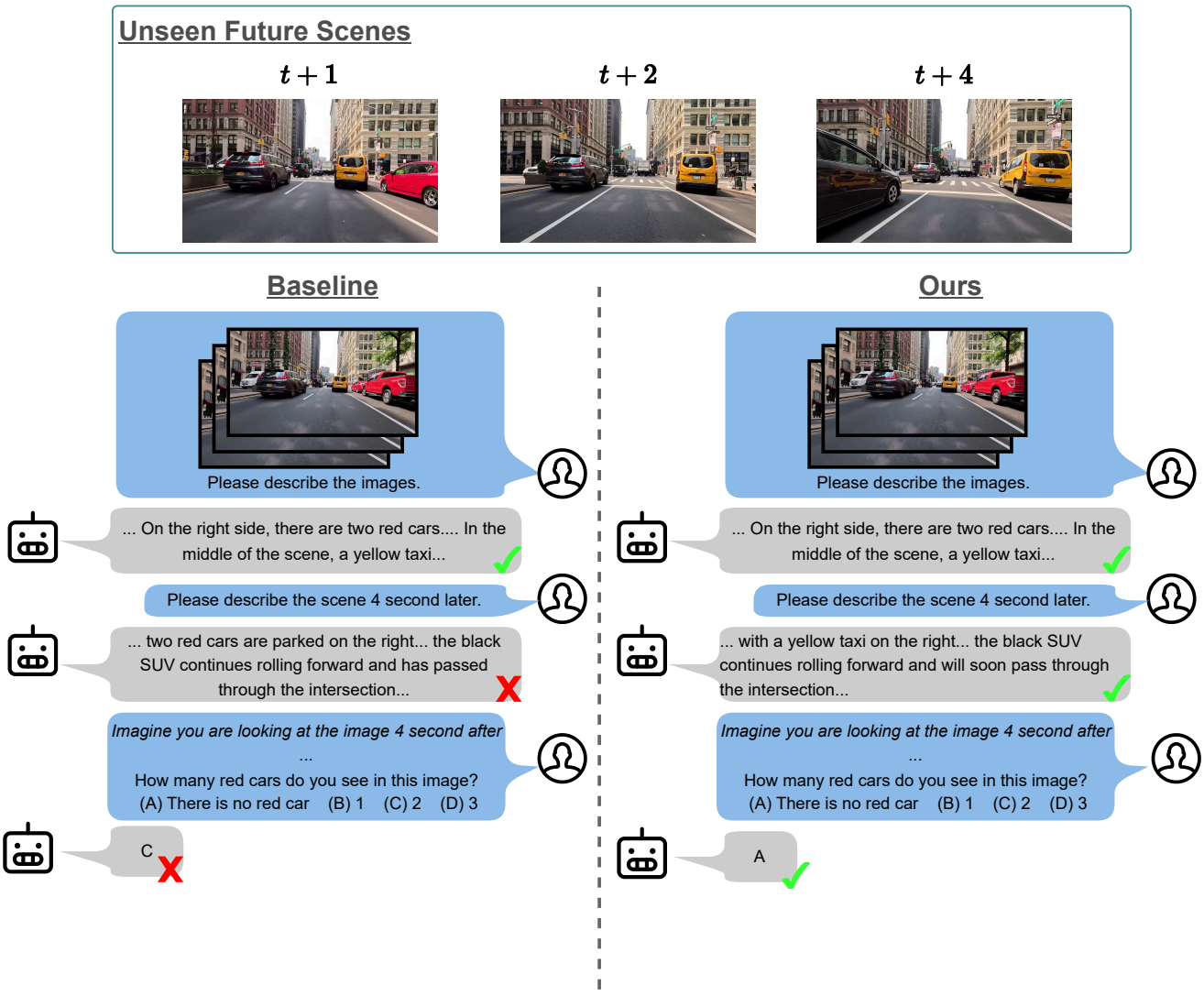


Figure 15. Quantitative result.

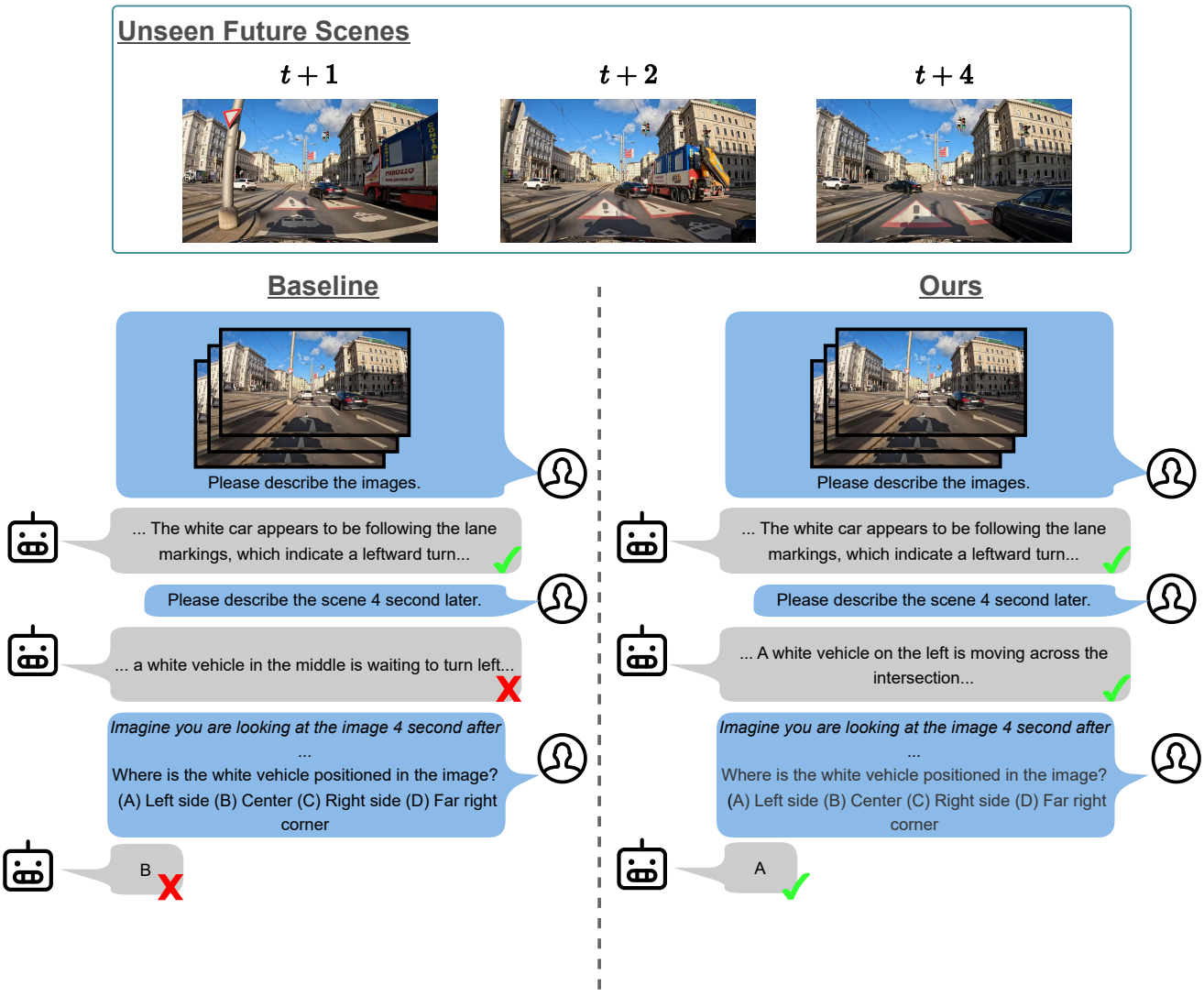


Figure 16. Quantitative result.

References

[1] Jihyun Janice Ahn and Wenpeng Yin. Prompt-reverse inconsistency: Llm self-inconsistency beyond generative randomness and prompt paraphrasing. *arXiv preprint arXiv:2504.01282*, 2025. 3

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: semantic propositional image caption evaluation. *CoRR*, abs/1607.08822, 2016. 4

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015. 2

[4] Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, et al. Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[5] 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. 1, 2, 6, 12

[6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6, 8

[7] 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*, pages 65–72, Ann Arbor, Michigan, 2005. Association for Computational Linguistics. 4

[8] Chun-Peng Chang, Shaoxiang Wang, Alain Pagani, and Didier Stricker. Mikasa: Multi-key-anchor & scene-aware transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14131–14140, 2024. 4

[9] Chun-Peng Chang, Alain Pagani, and Didier Stricker. 3d spatial understanding in mllms: Disambiguation and evaluation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13537–13544. IEEE, 2025. 4

[10] Chun-Peng Chang, Chen-Yu Wang, Julian Schmidt, Holger Caesar, and Alain Pagani. Seeing clearly, forgetting deeply: Revisiting fine-tuned video generators for driving simulation. *arXiv preprint arXiv:2508.16512*, 2025. 1

[11] Kai Chen, Yanze Li, Wenhua Zhang, Yanxin Liu, Pengxiang Li, Ruiyuan Gao, Lanqing Hong, Meng Tian, Xinhai Zhao, Zhenguo Li, et al. Automated evaluation of large vision-language models on self-driving corner cases. *arXiv preprint arXiv:2404.10595*, 2024. 2

[12] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14093–14100. IEEE, 2024. 2

[13] Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie-Francine Moens. Talk2car: Taking control of your self-driving car. *arXiv preprint arXiv:1909.10838*, 2019. 2

[14] Bahare Fatemi et al. Test of time: A benchmark for evaluating llms on temporal reasoning. *arXiv*, 2024. 1

[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. 13

[16] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 2, 13

[17] Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Dingkang Liang, Chong Zhang, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. *arXiv preprint arXiv:2503.19755*, 2025. 1

[18] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*, 2023. 4

[19] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2

[20] Akshay Gopalkrishnan, Ross Greer, and Mohan Trivedi. Multi-frame, lightweight & efficient vision-language models for question answering in autonomous driving. *arXiv preprint arXiv:2403.19838*, 2024. 2

[21] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[22] Mariam Hassan, Sebastian Stapf, Ahmad Rahimi, Pedro Rezende, Yasaman Haghighi, David Brüggemann, Isinsu Katircioglu, Lin Zhang, Xiaoran Chen, Suman Saha, et al. Gem: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control. *arXiv preprint arXiv:2412.11198*, 2024. 2

[23] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

[24] Xiaotao Hu, Wei Yin, Mingkai Jia, Junyuan Deng, Xiaoyang Guo, Qian Zhang, Xiaoxiao Long, and Ping Tan. Drivingworld: Constructingworld model for autonomous driving via video gpt. *arXiv preprint arXiv:2412.19505*, 2024. 2

[25] Zanming Huang, Jimuyang Zhang, and Eshed Ohn-Bar. Neural volumetric world models for autonomous driving. In

European Conference on Computer Vision, pages 195–213. Springer, 2024. 2

[26] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6, 8

[27] Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, et al. Emma: End-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2410.23262*, 2024. 1

[28] Fan Jia, Weixin Mao, Yingfei Liu, Yucheng Zhao, Yuqing Wen, Chi Zhang, Xiangyu Zhang, and Tiancai Wang. Adriver-i: A general world model for autonomous driving. *arXiv preprint arXiv:2311.13549*, 2023. 2

[29] Bo Jiang, Shaoyu Chen, Bencheng Liao, Xingyu Zhang, Wei Yin, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Senna: Bridging large vision-language models and end-to-end autonomous driving. *arXiv preprint arXiv:2410.22313*, 2024. 1

[30] Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025. 2

[31] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 4999–5007, 2017. 2

[32] Tarasha Khurana, Peiyun Hu, David Held, and Deva Ramanan. Point cloud forecasting as a proxy for 4d occupancy forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1116–1124, 2023. 2

[33] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision (ECCV)*, pages 563–578, 2018. 13

[34] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018. 2

[35] Guibiao Liao, Jiankun Li, and Xiaoqing Ye. Vlm2scene: Self-supervised image-text-lidar learning with foundation models for autonomous driving scene understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3351–3359, 2024. 2

[36] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. 4

[37] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv:2310.03744*, 2023. 1, 2, 14

[38] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023. 2, 6, 17, 18

[39] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 1, 2, 6, 7, 14, 18

[40] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*, 2023. 2, 13

[41] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023. 4

[42] Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. Drama: Joint risk localization and captioning in driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1043–1052, 2023. 2, 12, 13

[43] Sivabalan Manivasagam, Shenlong Wang, Kelvin Wong, Wenyuan Zeng, Mikita Sazanovich, Shuhan Tan, Bin Yang, Wei-Chiu Ma, and Raquel Urtasun. Lidarsim: Realistic lidar simulation by leveraging the real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11167–11176, 2020. 2

[44] Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. In *European Conference on Computer Vision*, pages 292–308. Springer, 2024. 2

[45] Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar, and Liu Ren. Vlp: Vision language planning for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14760–14769, 2024. 2

[46] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. 4

[47] 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, 12, 13

[48] 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 *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 6

[49] Katrin Renz, Long Chen, Elahe Arani, and Oleg Sinavski. Simlingo: Vision-only closed-loop autonomous driving with language-action alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11993–12003, 2025. 1

[50] Enna Sachdeva, Nakul Agarwal, Suhas Chundi, Sean Roelofs, Jiachen Li, Mykel Kochenderfer, Chiho Choi, and

Behzad Dariush. Rank2tell: A multimodal driving dataset for joint importance ranking and reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7513–7522, 2024. 2, 13

[51] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*, 2020. 4

[52] 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. 2, 4, 12, 13

[53] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Object referring in videos with language and human gaze. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4129–4138, 2018. 2

[54] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *CoRR*, abs/1411.5726, 2014. 4

[55] Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M. Alvarez. OmniDrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning. *arXiv:2405.01533*, 2024. 2, 12, 13

[56] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2023. 1, 2, 6, 7, 14

[57] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European Conference on Computer Vision*, pages 55–72. Springer, 2024. 2

[58] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14749–14759, 2024. 2

[59] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 5

[60] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*, 2024. 2, 13

[61] Rongwu Xu et al. Knowledge conflicts for llms: A survey. *arXiv*, 2024. 1

[62] Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, Jun Zhang, Andreas Geiger, Yu Qiao, and Hongyang Li. Generalized predictive model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 6, 12

[63] Zetong Yang, Li Chen, Yanan Sun, and Hongyang Li. Visual point cloud forecasting enables scalable autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14673–14684, 2024. 2

[64] Junwei You, Haotian Shi, Zhuoyu Jiang, Zilin Huang, Rui Gan, Keshu Wu, Xi Cheng, Xiaopeng Li, and Bin Ran. V2x-vlm: End-to-end v2x cooperative autonomous driving through large vision-language models. *arXiv preprint arXiv:2408.09251*, 2024. 2

[65] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024. 1, 2, 6, 14

[66] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277, 2021. 4

[67] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 6

[68] Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 1, 2

[69] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and Yang: Balancing and answering binary visual questions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[70] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 2, 6

[71] Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. *arXiv preprint arXiv:2403.06845*, 2024. 2

[72] 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. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023. 4

[73] Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao, and Alois C Knoll. Vision language models in autonomous driving: A survey and outlook. *IEEE Transactions on Intelligent Vehicles*, 2024. 2

[74] Zijian Zhou, Shikun Liu, Xiao Han, Haozhe Liu, Kam Woh Ng, Tian Xie, Yuren Cong, Hang Li, Mengmeng Xu, Juan-Manuel Pérez-Rúa, et al. Learning flow fields in attention for controllable person image generation. *arXiv preprint arXiv:2412.08486*, 2024. 2