Supplementary Material for NuPlanQA: A Large-Scale Dataset and Benchmark for Multi-View Driving Scene Understanding in Multi-Modal Large Language Models

Sung-Yeon Park[1], Can Cui[1], Yunsheng Ma[1],
Ahmadreza Moradipari[2], Rohit Gupta[2], Kyungtae Han[2], and Ziran Wang[1]

[1]Purdue University [2]Toyota InfoTech Labs
{sungyeon, ziran}@purdue.edu

1. Related Works

1.1. Video Understanding with Large Language Models

With the development of large language models (LLMs) in the image modality, video LLMs have also been actively explored. Most video LLMs adopt architectures similar to image LLMs, incorporating multi-modal adapters to capture temporal information by aggregating frame-level representations from videos [12, 15, 31]. Common approaches to video understanding in multi-modal LLMs (MLLMs) follow the LLaVA-style variation, where models are adapted to handle video inputs through lightweight projection layer [13, 16, 33]. For example, Video-LLaVA [12] encodes both image and video features into textual representations using LanguageBind [32], while a shared projection layer connected with the LLM. In addition to MLP-based or linear projection layers [7, 13, 14], vision-language adapters with cross-attention layers [1, 6, 30] are also widely used. Similarly, Video-LLaMA [31] utilizes a Q-former architecture [11] with learnable queries to bridge the gap between video and language representations. Beyond video understanding, it also integrates audio processing by training a separate Q-former branch for audio understanding. To address the limitations of video LLMs, such as handling diverse image resolutions, balancing performance trade-offs with image understanding, and improving long-term video comprehension, ongoing research aims to enhance the multi-modal capabilities of LLMs [10, 22, 25, 26, 29].

1.2. Visual Question Answering in Autonomous Driving

Early-stage research on autonomous driving with natural language began with video captioning. Initial works, such as BDD-X [9] and HDD [21], focused on describing drivers' actions and the reasoning behind them in specific driving scenarios. Building on these datasets, captioning tasks in driving scenes have been widely explored, as seen in ADAPT [8]. With the advent of LLMs, research in autonomous driving extended beyond captioning to visual question answering (VQA), enabling more flexible and instruction-driven interactions. Early efforts in this direction, such as DriveGPT4 [28] and VLAAD [19], leveraged BDD-X and HDD to cost-effectively generate question-answer pairs using ChatGPT. To introduce object-level logical dependencies into QA pairs, DriveLM [23] and NuScenesQA [20] utilized the nuScenes dataset [3], which offers more detailed annotations, including 3D bounding boxes and object tracks.

Following these works, several datasets based on nuScenes [3], such as NuPrompt [27] and NuInstruct [5], were proposed. However, the heavy reliance on nuScenes limits the diversity of training data, restricting MLLMs from learning a broader range of driving scenarios. To address this limitation, LingoQA [17] introduced free-form QA using newly collected videos from vehicles. Nevertheless, since it only provides front-view images, it fails to capture surrounding environmental context. Despite active research in this domain, the size of available datasets and the variety of included driving scenarios remain limited. Moreover, a key limitation of existing studies is their reliance on n-gram precision metrics, such as BLEU [18], METEOR [2], and CIDEr [24], to evaluate MLLMs. These metrics often fail to penalize factual errors, such as incorrect turn directions or misidentified traffic lights, as they primarily assess lexical similarity. As a result, models may receive high scores even when generating incorrect responses. Furthermore, in tasks that do not involve explicit visual grounding or motion planning, there is still a lack of standardized benchmarks for visual question answering, making fair comparison across models challenging.

Task	**NuPlanQA-1M**	**NuPlanQA-Eval***
Traffic Light	88,836 (9.1%)	203 (11.3%)
Weather/Lighting	88,874 (9.1%)	217 (12.0%)
Road Type/Condition	177,720 (18.2%)	192 (10.7%)
Surrounding Objects	88,870 (9.1%)	181 (10.0%)
Traffic Flow	88,853 (9.1%)	196 (10.9%)
Key Objects	88,875 (9.1%)	220 (12.2%)
Ego-vehicle Maneuver	177,721 (18.2%)	191 (10.6%)
Situation Assessment	88,864 (9.1%)	202 (11.2%)
Action Recommendation	88,782 (9.1%)	199 (11.0%)
Total	977,395 (100%)	1,801 (100%)

Table 1. **Dataset distribution per task.** * indicates test set of NuPlanQA-Eval.

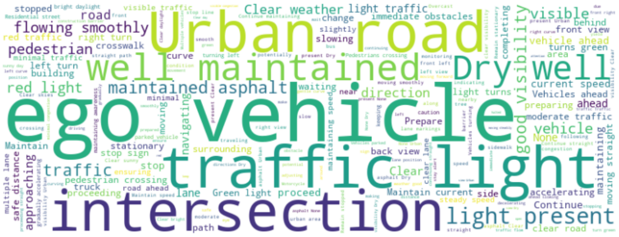


Figure 1. **Word cloud of NuPlanQA-1M.**

2. Dataset

In this section, we provide a detailed overview of the generation process, distribution, and examples of NuPlanQA-1M and NuPlanQA-Eval.

2.1. Dataset Generation

NuPlanQA-1M. To generate a large-scale, high-quality dataset of QA pairs in a scalable manner—incorporating both multi-timestep and multi-view images—we utilize GPT-4o (gpt-4o-2024-08-06). To ensure precise responses from GPT-4o, we structure data from nuPlan [4] into per-frame velocities, steering angles, and textual descriptions of traffic situations. In particular, the traffic situation information plays a crucial role in providing a detailed and comprehensive understanding of the scenes. We sample frames from 1.5-second video segments, and using this structured information, GPT-4o generate responses. The specific prompt used for this generation process is depicted in Figure 2. Through experimental evaluations of GPT-4o outputs, we found that it performs best when prompted to answer in a chain-of-thought manner, progressing from low-level perception tasks to high-level reasoning tasks. Based on this insight, we design subtasks ranging from traffic light detection to vehicle maneuver recommendation, resulting in 11 initial subtasks. By structuring responses in this step-by-step manner, we obtain more accurate scene interpretations. In the final NuPlanQA-1M dataset, these subtasks are refined and consolidated into a total of nine subtasks. After we generate QA pairs with GPT-4o, to guarantee its quality, we utilize several filtering criteria. For the traffic light detection task, we use the traffic light status provided by the original nuPlan dataset. Since multiple signals can be active simultaneously, we leverage both GPT-4o responses and raw metadata to resolve ambiguity. Additionally, for tasks involving vehicle behavior, we filter out samples where GPT-4o responses are inconsistent with control signals such as steering angle and velocity.

NuPlanQA-Eval. After generating QA pairs, we split the dataset into training and evaluation sets, ensuring careful scenario separation to prevent any overlap between them. For the evaluation dataset, we convert the QA pairs into a multiple-choice QA format by generating three additional answer options—alongside the correct answer—using GPT-4V. The prompt used for this task is shown in Figure 3. However, since the generated choices might sometimes be too easy to eliminate, contain overlapping options, or even result in an incorrect true answer, we manually review and refine the samples to ensure quality. To support zero-shot and few-shot inference, as well as fine-tuning for evaluation, we further split the evaluation dataset into train, validation, and test sets. Only the test set is used for final evaluation. As illustrated in Figure 4, to prevent models from directly referring to traffic situation information, we exclude it from the evaluation dataset. Our evaluation set includes velocity and steering angle data for a 1.5-second window; however, considering the required length of historical information for inference, this can be adjusted using raw data from nuPlan.

2.2. Dataset Statistics

The number and proportion of each subtask in the dataset are shown in Table 1. In NuPlanQA-1M, each subtask accounts for 9.1% of the dataset, except for road type/condition and ego-vehicle maneuver tasks, which have double the proportion. However, the test set of NuPlanQA-Eval maintains a relatively even distribution across all tasks to ensure fair model evaluation. The train and validation sets of NuPlanQA-Eval follow the same distribution as NuPlanQA-1M. As a result, NuPlanQA-1M contains 977,395 QA pairs and NuPlanQA-Eval consists of 1,801 QA pairs. As shown in Figure 1, our dataset predominantly features terms such as “ego-vehicle”, “traffic light”, “intersection”, and “urban road”. Considering the complexity of urban scenarios, we primarily include scenes from urban environments, excluding some scenarios from Singapore.

2.3. Dataset Examples

We present additional example QAs for each subtask in NuPlanQA-1M in Table 2. Detailed descriptions for each task are as follows:

```
Prompt = '''You are driving from inside the vehicle cabin. The image shows views from the ego-vehicle:
front, front right/left, back right/left and back, arranged from top to bottom.
The image captures past 1.5 seconds, with three frames extracted at regular intervals, progressing from
top to bottom.
Below is additional information from your vehicle, extracted 5 times over past 1.5 seconds. Use changes
in velocity and steering angle to determine if the vehicle is slowing, accelerating, turning, curving,
changing lanes, or adjusting. Negative steering angles indicate a right turn and positive indicate a
left turn.

[Given Information]
 • Ego-Vehicle Lane Traffic Situation:
['following_lane_without_lead','high_magnitude_speed','following_lane_without_lead','following_lane_wit
hout_lead', '']
 • Velocity(m/s): [11.959, 11.964, 11.961, 12.059, 12.271]
 • Steering angles: [0.061, -0.092, -0.128, 0.134, 0.181]

Follow the same template as below, however, never repeat the same sentence. Answer based on what you
observed. Answer for each categories within a sentence except the key objects. Don't mention
\"frames\", \"car\". Indicate your vehicle as ego-vehicle.\n\n

- Example
{\n
  a. Traffic Light: [No traffic light present. Red/Green/Yellow lights are visible.]\n
  b. Visible Objects: Vehicles turning left at the intersection. Trucks at the front left view.
     Pedestrians at the front right view. None.\n
  c. Key Objects: (Important to notice for safe driving) The vehicle ahead slowing down, potentially
     about to make a turn. Vehicles are moving through at the intersection.\n
  d. Road Type: Urban road with heavy traffic. Residential street, narrow with parked vehicles on the
     sides. Highway with curved path.\n
  e. Weather/Lighting Conditions: Rainy, reduced visibility due to light rain.\n
  f. Road Conditions: Dry, well-maintained asphalt. Wet roads due to recent rain, with a few puddles
     forming near the curbs. \n
  g. Traffic Flow: Traffic is flowing smoothly in both directions. There is moderate congestion ahead,
     likely due to construction work. \n
  h. Controls: Since angles rising to 12.44 which is more than 7 and then decreasing, the ego-vehicle
     straightens after making left-turn. Since the velocity is near 0 and making a sharp turn, the ego-
     vehicle seems exiting from the parking space.\n
  i. Current Vehicle Maneuver: The ego-vehicle is completing a left turn while gradually accelerating.
     The ego-vehicle is changing lane into left to avoid traffic ahead. \n
  j. Situation Assessment: The ego-vehicle is approaching an intersection on left turning lane where the
     traffic light is green.\n
  k. Vehicle Maneuver Recommendation: Maintain speed and continue straight through the green light,
     while keeping an eye on the pedestrian. Keep safe distance with the vehicle behind on the left lane
     before merging.\n
}\n

Response in json format having each categories as keys.'''
```

Figure 2. **Prompt for generating NuPlanQA-1M.** By providing control parameters for past frames and detailed textual descriptions of traffic situations, GPT-4o achieves a better understanding of scenes. Through chain-of-thought generation with subtasks, it gains a deeper comprehension of the context.

- ***Traffic Light***: Detecting the presence and color of traffic lights that the ego-vehicle should obey or be aware of.
- ***Weather/Lighting Conditions***: Identifying weather and lighting conditions that affect visibility on the road.
- ***Road Type/Conditions***: Recognizing the clarity of lane markings, the type of road, and conditions.
- ***Surrounding Objects***: Detecting objects around the ego-vehicle that the driver should be aware of.
- ***Traffic Flow***: Analyzing the current traffic flow, including congestion levels and whether vehicles are moving or stationary.
- ***Key Objects***: Identifying critical objects that are essential for the ego-vehicle's safe navigation and planning.

```
Prompt = '''Generate three different responses that can be choices with given question. Since given
answer is true for the question, you should generate three more choices which cannot be an true
answer.\n For the question asking about traffic lights, choices can be presence/absence or different
colors from the true answer. Three choices should be different from each other.\n
Response in json format.

[Example]\n
- Question: Is the traffic moving smoothly?\n
- Answer: Traffic is flowing smoothly in both directions.\n

{\"choices\": [
    "Traffic is heavily congested in both directions.",
    "Roadblocks are causing significant delays.",
    "There is a major accident causing traffic to slow down."
]}'''
```

Figure 3. **Prompt for multiple-choice option generation.** Possible answer sets are generated using GPT-4o and further refined by human annotators.

```
Prompt = '''You are driving on the road. Based on given images of the surroundings of your vehicle and
information from your vehicle extracted past 1.5 seconds, choose right answer among (a) to (d) for the
given question.\n

- Velocity(m/s): [10.391, 10.257, 10.145, 10.041, 9.811]\n
- Steering angles: [-0.149, -0.024, 0.042, 0.047, 0.068]\n

Question:  What is the situation the vehicle currently in?\n

(a) The ego-vehicle is navigating through heavy rain and poor visibility.\n
(b) The ego-vehicle is driving on a road with multiple construction zones and detours.\n
(c) The ego-vehicle is approaching an intersection with no immediate obstructions.\n
(d) The ego-vehicle is stuck in a traffic jam and unable to move.'''
```

Figure 4. **Prompt for evaluation.** Multiple-choice options are provided along with control parameters for evaluation.

- ***Ego-vehicle Maneuver***: Determining the movement of the ego-vehicle based on control signals and visual information.
- ***Situation Assessment***: Providing a comprehensive description of the current driving scenario based on gathered information.
- ***Action Recommendation***: Recommending safe driving actions for the ego vehicle by considering both internal and external conditions.

3. Experiments

3.1. Ablation study

Table 3 presents the ablation study results for each subtask. Compared to other configurations, the BEV-Fusion model with multi-view, multi-frame inputs achieves the highest scores in most subtasks. However, the BEV-Fusion model with single-view inputs performs better on the traffic light and weather/lighting condition tasks, both of which fall under the road environment perception skill category. This suggests that for static object detection, single-view inputs contribute more effectively to perception. In particular, traffic lights, which are typically positioned in the front-view image, can be detected more accurately with single-view inputs by focusing on the most relevant image where a traffic light is most likely to appear. For all other subtasks—except for traffic light and weather/lighting condition—the BEV-Fusion module with multi-view, multi-frame inputs outperforms other settings, including achieving the highest average score.

To evaluate the effectiveness of BEV features alone, we compare different input configurations in Table 4. For the BEV-only setup, BEV features from the encoder are passed directly to the multi-modal projector, and the model is trained for one epoch on NuPlanQA-1M. We also assess the BEV-fusion setting by providing black frames to the im-

Task	**Example**
Traffic Light	**Q1**: Is there a traffic signal the ego-vehicle should give priority to? **A1**: No traffic light present. **Q2**: Are there any traffic signals the ego-vehicle should obey? **A2**: Red lights are visible.
Weather/Lighting Conditions	**Q1**: What are the current weather and lighting conditions? **A1**: Clear and sunny. **Q2**: Are there any weather conditions affecting the driving environment? **A2**: Overcast, with adequate visibility.
Road Type/Conditions	**Q1**: What is the condition of the road surface? **A1**: Dry, well-maintained asphalt. **Q2**: How would you describe the current road type? **A2**: Urban road with intersections.
Surrounding Objects	**Q1**: Which objects are present in the scene? **A1**: Vehicles stopped ahead in both directions. Buildings along the street. **Q2**: Which objects are present in the scene? **A2**: Vehicles entering the intersection. Vehicles visible behind the ego-vehicle.
Traffic Flow	**Q1**: Is the traffic moving smoothly? **A1**: Traffic is light with vehicles moving in both directions. **Q2**: What is the current state of the traffic flow? **A2**: Traffic is minimal with stationary vehicles ahead.
Key Objects	**Q1**: Are there any significant objects ahead that the driver needs to be aware of? **A1**: None currently affecting driving. **Q2**: What is the most important object nearby that the driver should focus on? **A2**: Large truck in front with ’oversize load’ sign.
Ego-vehicle Maneuver	**Q1**: What is the vehicle’s current movement? **A1**: The ego-vehicle is traveling straight while slightly decelerating. **Q2**: What is the vehicle’s current maneuver? **A2**: The ego-vehicle is following the lane while slightly curving to the right.
Situation Assessment	**Q1**: What is the overall situation assessment? **A1**: The ego-vehicle is at an intersection waiting for the light to change. **Q2**: How would you assess the current driving situation? **A2**: The ego-vehicle is approaching an intersection with pedestrians crossing.
Action Recommendation	**Q1**: What action is advised for maintaining safety? **A1**: Continue turning left while monitoring for any obstacles or vehicles entering the intersection from other directions. **Q2**: What is the recommended maneuver for the vehicle? **A2**: Maintain a steady speed while ensuring a safe gap with the vehicle ahead.

Table 2. **Examples of NuPlanQA-1M.** NuPlanQA-1M contains a diverse set of questions and detailed responses.

age encoder, disabling visual input while preserving the fusion pathway. As shown, BEV features alone yield limited performance. While extended training may improve results, these findings underscore that BEV features are most effective when fused with image features, supporting our fusion-based design.

3.2. Qualitative Results

In Figure 6, we present qualitative results on NuPlanQA-Eval across nine subtasks. By comparing with LLaVA-OneVision-7B [10], we highlight cases where MLLMs fail to make correct predictions. Errors in traffic light perception, vehicle maneuvers, surrounding objects, and other road information lead to a lack of overall understanding of traffic situations. In contrast, BEV-LLM demonstrates superior performance by accurately interpreting these elements.

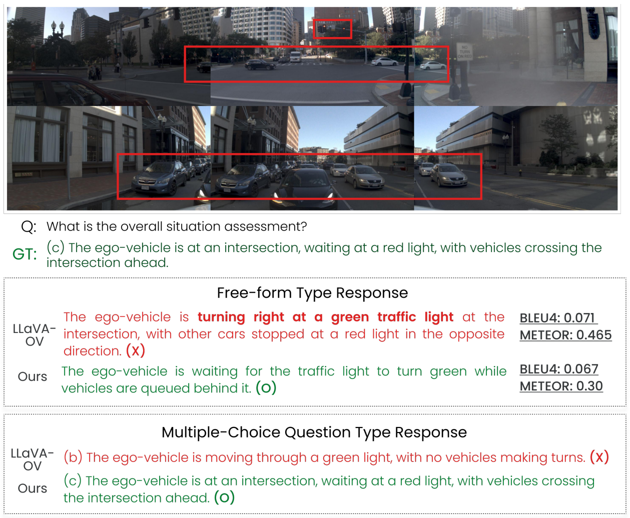
Figure 5 presents an example case where common text-based metrics fail. Since n-gram-based metrics such as BLEU-4, METEOR, and CIDEr evaluate generated sentences based on n-gram matching, they struggle to capture the context of road scenes and fail to recognize important cues expressed with different phrases. For instance, in Figure 5, despite the LLaVA-OneVision [10] generating incorrect answers, BLEU-4 and METEOR assign it a higher score than the correct response from BEV-LLM. However, using a multiple-choice question format, the evaluation successfully identifies which model generates the correct answer. Although alternative metrics like SPICE, which rely on scene graphs, exist, they also face limitations in han-

Method	Input	Road Env. Perception			Spatial Relations Recog.			Ego-centric Reasoning			Total
		Trfc. Light	*Wea -ther*	*Road Type*	*Sur. Obj.*	*Trfc. Flow*	*Key Obj.*	*Ego Ctrl.*	*Situ. Asse.*	*Act. Rec.*	
Baseline	MV + MF	52.2	81.6	72.4	65.2	62.2	55.0	62.8	69.8	68.3	65.5
+BEV-Fusion	SV + SF	<u>62.1</u>	**90.8**	74.0	69.1	68.4	60.5	68.1	75.7	73.9	71.4
+BEV-Fusion	SV + MF	**63.5**	84.9	78.6	72.4	69.9	61.8	72.8	76.2	71.9	72.4
+BEV-Fusion	MV + SF	58.1	<u>87.1</u>	<u>84.9</u>	<u>75.1</u>	<u>75.0</u>	<u>66.8</u>	<u>76.0</u>	<u>81.2</u>	<u>81.4</u>	<u>76.2</u>
+BEV-Fusion	MV + MF	61.1	<u>89.4</u>	**89.6**	**78.5**	**75.5**	**68.2**	**79.1**	**83.2**	**83.4**	**78.7**

Table 3. **Ablation study results on BEV-LLM across nine subtasks.** The metric used is accuracy, with detailed experimental results illustrated for each subtask. MV/SV denotes multi-view/single-view, and MF/SF denotes multi-frame/single-frame inputs. The best-performing model in each task is **bolded**, while the second-best is <u>underlined</u>.

Input	Road Env.	Spatial Rel.	Ego Rea.	Avg.
Image + BEV	80.0	74.1	81.9	78.7
Image features	68.7	60.8	67.0	65.5
BEV-fusion features	24.9	16.9	15.1	19.0
BEV-encoder features	23.7	10.3	14.8	16.3

Table 4. **Ablation study results on BEV-LLM across different input features.** The metric used is accuracy. Multi-view and multi-frame inputs are used.



Q: What is the overall situation assessment?

GT: (c) The ego-vehicle is at an intersection, waiting at a red light, with vehicles crossing the intersection ahead.

Free-form Type Response

LLaVA-OV	The ego-vehicle is **turning right at a green traffic light** at the intersection, with other cars stopped at a red light in the opposite direction. **(X)**	<u>BLEU4: 0.071</u> <u>METEOR: 0.465</u>
Ours	The ego-vehicle is waiting for the traffic light to turn green while vehicles are queued behind it. **(O)**	<u>BLEU4: 0.067</u> <u>METEOR: 0.30</u>

Multiple-Choice Question Type Response

LLaVA-OV	(b) The ego-vehicle is moving through a green light, with no vehicles making turns. **(X)**
Ours	(c) The ego-vehicle is at an intersection, waiting at a red light, with vehicles crossing the intersection ahead. **(O)**

Figure 5. **Comparison of free-form and multiple-choice question responses.** BLEU-4 and METEOR are calculated for free-form responses. LLaVA-OV refers to the LLaVA-OneVision model.

dling ambiguities in scene representation and spatial relationships.

References

[1] 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] 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. 1

[3] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[4] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. Nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles, 2022. 2

[5] Xinpeng Ding, Jianhua Han, Hang Xu, Xiaodan Liang, Wei Zhang, and Xiaomeng Li. Holistic autonomous driving understanding by bird'view injected multi-modal large models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13668–13677, 2024. 1

[6] Aaron Grattafiori et al. The llama 3 herd of models, 2024. 1

[7] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Shaoqi Dong, Xiong Wang, Di Yin, Long Ma, et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024. 1

[8] Bu Jin, Xinyu Liu, Yupeng Zheng, Pengfei Li, Hao Zhao, Tong Zhang, Yuhang Zheng, Guyue Zhou, and Jingjing Liu. Adapt: Action-aware driving caption transformer. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7554–7561, 2023. 1

[9] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1

[10] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. 1, 5

[11] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with

frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 1

[12] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 1

[13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1

[14] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Vishwesh Nath, Jinyi Hu, Sifei Liu, Ranjay Krishna, Daguang Xu, Xiaolong Wang, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, and Yao Lu. Nvila: Efficient frontier visual language models, 2024. 1

[15] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Da Li, Pengcheng Lu, Tao Wang, Linmei Hu, Minghui Qiu, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability, 2023. 1

[16] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024. 1

[17] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, Elahe Arani, and Oleg Sinavski. Lingoqa: Visual question answering for autonomous driving, 2024. 1

[18] 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 on Association for Computational Linguistics*, page 311–318, USA, 2002. Association for Computational Linguistics. 1

[19] SungYeon Park, MinJae Lee, JiHyuk Kang, Hahyeon Choi, Yoonah Park, Juhwan Cho, Adam Lee, and DongKyu Kim. Vlaad: Vision and language assistant for autonomous driving. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 980–987, 2024. 1

[20] 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. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2025. 1

[21] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7699–7707, 2018. 1

[22] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024. 1

[23] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering, 2024. 1

[24] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015. 1

[25] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1

[26] Xidong Wang, Dingjie Song, Shunian Chen, Chen Zhang, and Benyou Wang. Longllava: Scaling multi-modal llms to 1000 images efficiently via hybrid architecture, 2024. 1

[27] Dongming Wu, Wencheng Han, Tiancai Wang, Yingfei Liu, Xiangyu Zhang, and Jianbing Shen. Language prompt for autonomous driving. *ArXiv*, abs/2309.04379, 2023. 1

[28] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee. K. Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model, 2024. 1

[29] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models, 2024. 1

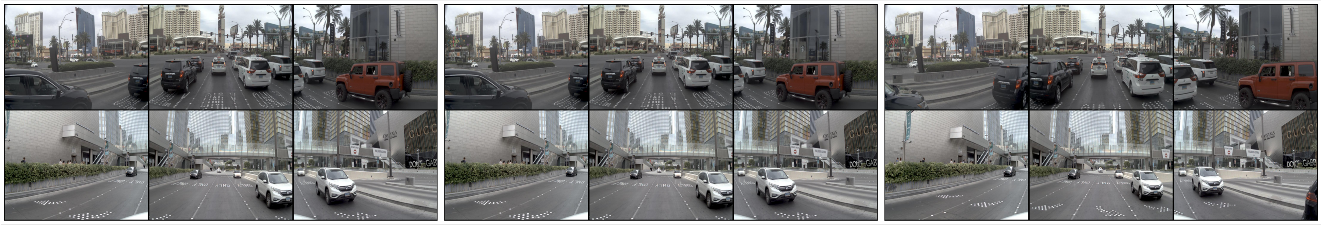
[30] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl: Modularization empowers large language models with multimodality, 2024. 1

[31] 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

[32] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Wang HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Wan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment, 2023. 1

[33] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1

(A) Traffic Light

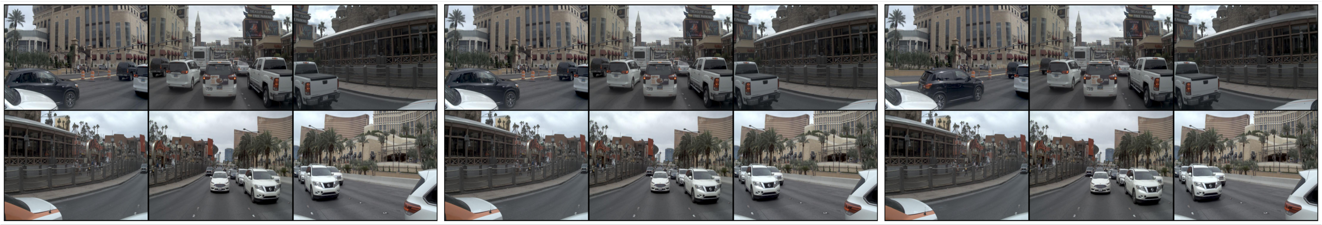


Question: Are traffic signals present in the environment?

BEV-LLM: (a) Red light is visible.

LLaVA-OV: (d) No traffic light present.

(B) Weather/Lighting Condition

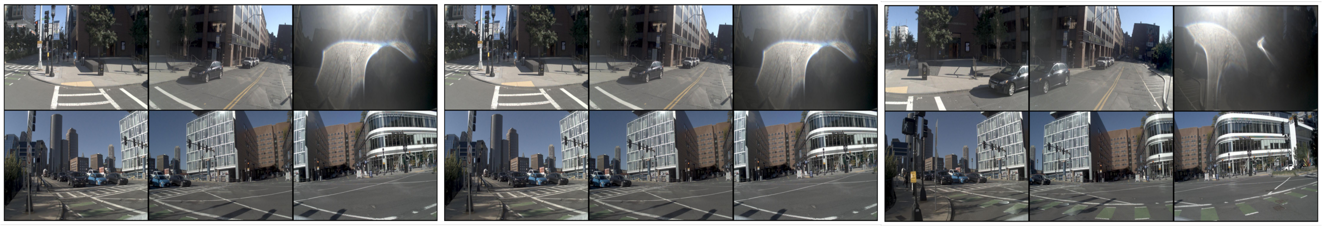


Question: What are the current weather and lighting conditions?

BEV-LLM: (c) Cloudy with moderate visibility.

LLaVA-OV: (d) Clear skies with excellent visibility.

(C) Road Type/Condition

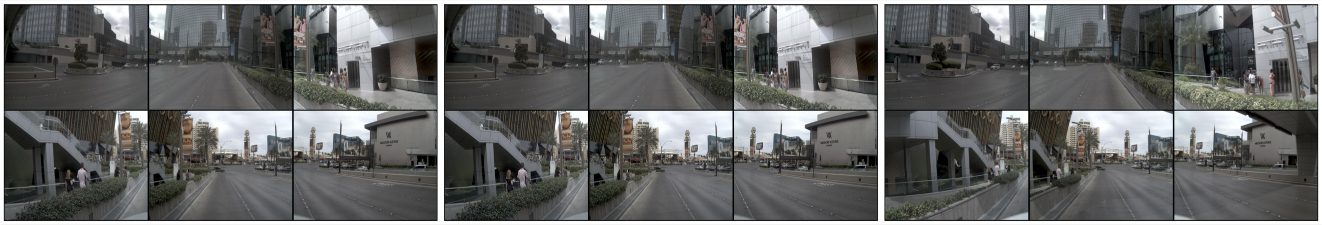


Question: What type of road is the vehicle currently on?

BEV-LLM: (d) Urban road with buildings on both sides.

LLaVA-OV: (c) Highway with multiple lanes and high speed limits.

(D) Surrounding Objects

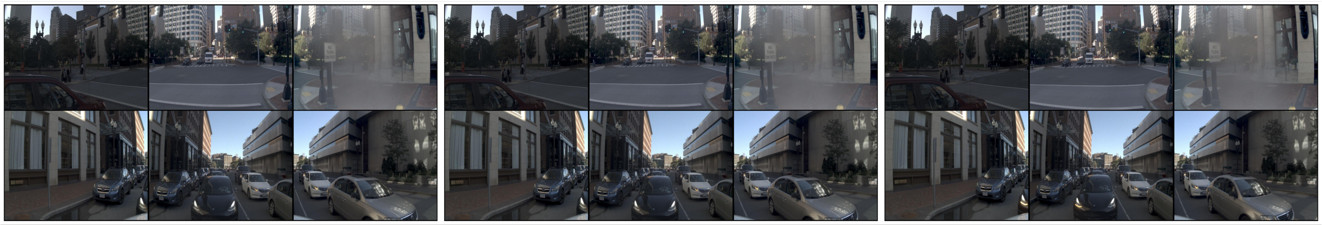


Question: Are there any visible objects around the vehicle?

BEV-LLM: (a) Pedestrians at the front right view.

LLaVA-OV: (b) There are no pedestrians or obstacles in sight.

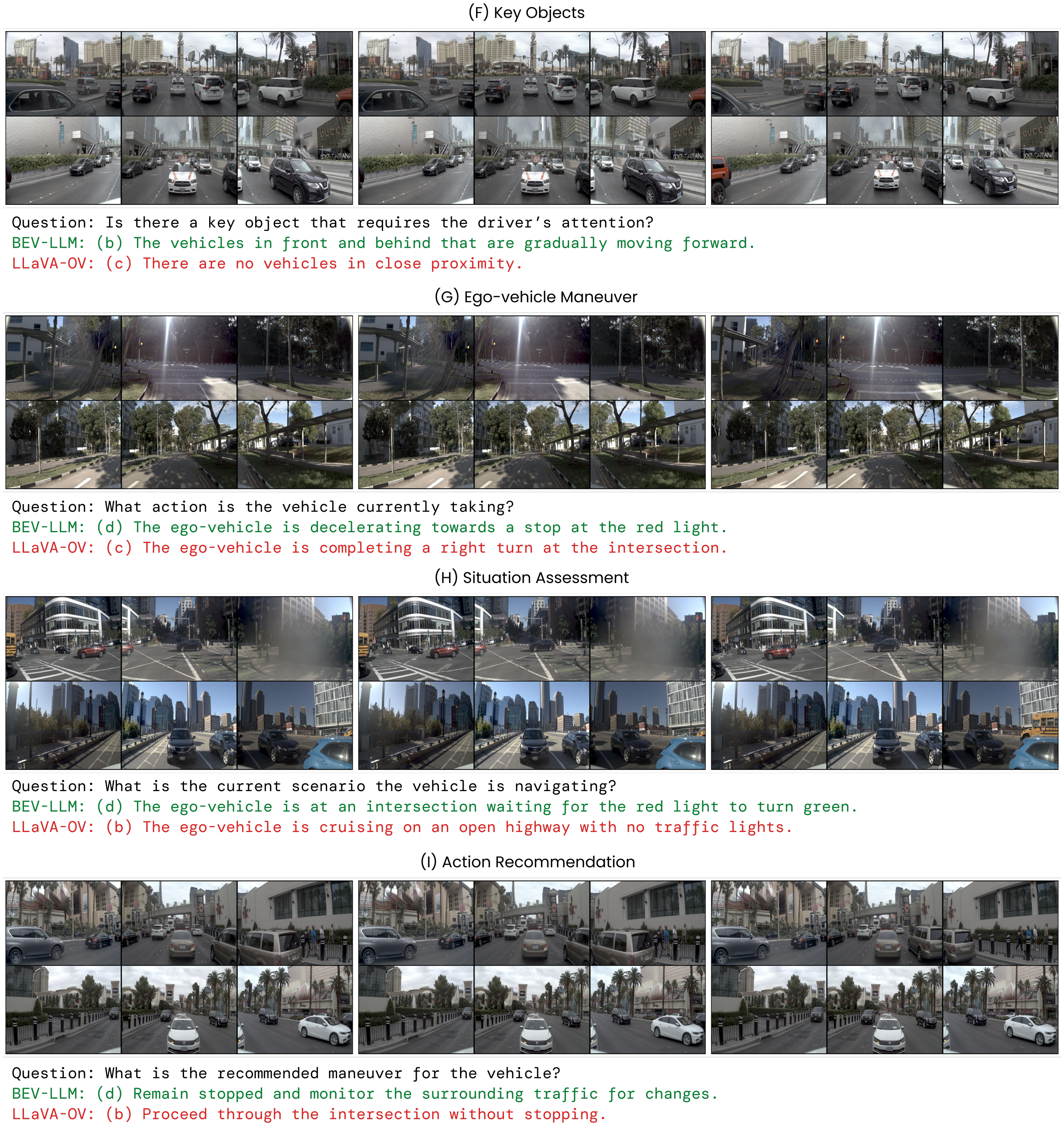
(E) Traffic Flow



Question: Is the traffic moving smoothly?

BEV-LLM: (d) Moderate traffic with stationary vehicles.

LLaVA-OV: (b) Traffic is flowing smoothly in both directions.



(F) Key Objects

Question: Is there a key object that requires the driver's attention?
BEV-LLM: (b) The vehicles in front and behind that are gradually moving forward.
LLaVA-OV: (c) There are no vehicles in close proximity.

(G) Ego-vehicle Maneuver

Question: What action is the vehicle currently taking?
BEV-LLM: (d) The ego-vehicle is decelerating towards a stop at the red light.
LLaVA-OV: (c) The ego-vehicle is completing a right turn at the intersection.

(H) Situation Assessment

Question: What is the current scenario the vehicle is navigating?
BEV-LLM: (d) The ego-vehicle is at an intersection waiting for the red light to turn green.
LLaVA-OV: (b) The ego-vehicle is cruising on an open highway with no traffic lights.

(I) Action Recommendation

Question: What is the recommended maneuver for the vehicle?
BEV-LLM: (d) Remain stopped and monitor the surrounding traffic for changes.
LLaVA-OV: (b) Proceed through the intersection without stopping.

Figure 6. **Qualitative results from evaluation.** Results from LLaVA-OneVision-7B and BEV-LLM are shown for comparison. False cases for LLaVA-OneVision (red) and correct cases for BEV-LLM (green) are illustrated to highlight scenarios where existing MLLMs fail. Three past timeframes used as inputs are presented from left to right.