Can MLLMs Reason over Surround-View Driving Videos? A Multi-View Autonomous Driving QA Benchmark

Anonymous ACL submission

Abstract

Surround-view video understanding in autonomous driving is a crucial capability for Multimodal Large Language Models (MLLMs) to achieve reliable driving reasoning. Real-world driving often requires information beyond the front-view camera. Existing MLLMs primarily focus on front-view information, while their spatio-temporal understanding and reasoning abilities remain unknown. To address this gap, we propose MVDriveBench, a high-information-density video QA benchmark for autonomous driving, characterized by synchronized surround-view inputs, spatio-temporal understanding, and reasoning-chain-oriented evaluation. We design 12 categories of QA tasks that cover the complete reasoning pipeline from low-level visual perception to high-level driving decision-making, enabling a systematic evaluation of MLLMs' surround-view spatio-temporal understanding and reasoning capabilities in complex driving environments. We conduct a comprehensive evaluation of 21 open-source, closed-source, and expert MLLMs on MVDriveBench. The results show that current MLLMs still exhibit insufficient surround-view spatio-temporal reasoning ability.

1 Introduction

In recent years, MLLMs have achieved remarkable progress in tasks such as image understanding, video question answering, visual reasoning, and open-ended dialogue (Yin et al., 2024). As model capabilities continue to improve, systematically and objectively evaluating the perception, reasoning, and decision-making abilities of MLLMs has become a central issue in multimodal research (Fu et al., 2024). To this end, researchers have proposed a wide range of general-purpose multimodal benchmarks (Li et al., 2024b; Fu et al., 2025; Wu et al., 2024; Mangalam et al., 2023; Liu et al., 2024) to evaluate MLLMs. These benchmarks

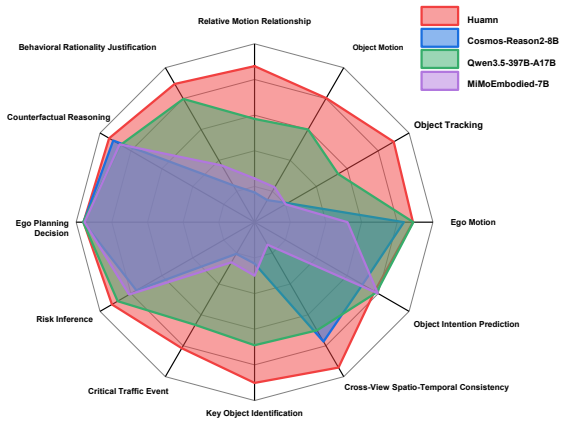


Figure 1: Performance of MLLMs on MVDriveBench.

have advanced the development of MLLMs on open-domain vision-language tasks. However, they typically focus on visual recognition and temporal understanding in general scenarios, and therefore cannot adequately capture the highly safety-critical, spatially constrained, and decision-oriented reasoning requirements of autonomous driving.

Autonomous driving is one of the key application scenarios for bringing MLLMs into the real physical world (Cui et al., 2024), and vision-language benchmarks for autonomous driving have therefore attracted increasing attention. Real-world driving often cannot be made based solely on the front-view camera. For example, assessing lane-change safety requires jointly considering the spatial relationships among vehicles in the front-side, rear-side, and rear views. In other words, many safety-critical questions in autonomous driving are inherently dependent on synchronized multi-view videos. However, existing autonomous driving benchmarks, which typically rely on single-view videos, discrete images, multi-view static images, or structured scene representations, are still insufficient for comprehensively evaluating the ability of MLLMs to understand and reason over surround-

Benchmark / Dataset	AD	#Videos/Images	#QA	#Tasks	Anno	Closed-ended	Open-ended	Multi-video	Multi-view
MVBench	✗	4K	4K	20	A	✓	✗	✓	✗
Video-MME	✗	900	2.7K	30	M	✓	✗	✓	✗
LongVideoBench	✗	3,763	6,678	17	M	✓	✗	✓	✗
EgoSchema	✗	5K	5K	1	M	✓	✗	✓	✗
TempCompass	✗	410	7,540	5	M+A	✓	✓	✓	✗
BDD-X	✓	6,984	–	2	M	✗	✓	✓	✗
NuScenes-QA	✓	34K	460K	10	M+A	✓	✗	✗	✓
LingoQA	✓	28K	419K	4	M	✗	✓	✓	✗
DriveLM	✓	4871	443K	5	M+A	✓	✓	✗	✓
DriveBench	✓	19.2K	20,498	4	M+A	✓	✓	✗	✗
Reason2Drive	✓	–	600K	3	A	✗	✓	✓	✗
Talk2BEV	✓	1K	20K	4	M	✓	✓	✗	✗
DriveLMM-o1	✓	539	4633	3	M+A	✓	✓	✗	✓
NuPlanQA	✓	90K	1M	9	A	✓	✗	✗	✓
STSBench	✓	43	971	2	M+A	✓	✗	✓	✓
MVDriveBench (Ours)	✓	**2418**	**4303**	**12**	**M+A**	✓	✓	✓	✓

Table 1: Comparison between MVDriveBench and existing video understanding and autonomous driving visual-language benchmarks or datasets. AD denotes Autonomous Driving. Anno denotes annotation type: M = manual annotation and A = automatic annotation.

view driving videos. This leaves a fundamental question insufficiently explored: ***Do current MLLMs truly possess the ability to understand and reason over surround-view driving videos?***

To address this question, we propose **MVDriveBench**. Specifically, MVDriveBench captures the surrounding traffic environment of the ego vehicle through multiple time-synchronized onboard cameras, requiring models to understand both spatial and temporal aspects of driving scenarios. In addition, each QA sample is accompanied by a corresponding reasoning rationale as an auxiliary reference. MVDriveBench is built with a semi-automatic pipeline emphasizing multi-view necessity and two-stage prompt generation. We select real-world driving scenarios that require information beyond the front-view camera, and convert six synchronized onboard views into a unified surround-view video input. QA pairs are then generated through a two-stage process: scene understanding, which analyzes object locations, motions, cross-view relations, and risks, followed by QA generation over 12 driving reasoning categories. Manual verification is finally applied to remove unsupported or erroneous samples, yielding MVDriveBench for model evaluation.

Overall, our evaluation of 21 MLLMs on MVDriveBench reveals that existing models still struggle with reliable surround-view driving video understanding: even the best-performing model, GPT-5, lags behind human performance by 9.8 points in *Overall* score, while Specialist Models achieve only 45.7–54.4. Meanwhile, as shown in Figure 1, the inference capability of the MoE open-source model with ultra-large parameters is also insufficient.

The main contributions of this paper are summarized as follows:

- We propose **MVDriveBench**, a multi-view QA benchmark for surround-view video understanding in autonomous driving. It fills the gap in existing benchmarks regarding surround-view video-based driving reasoning.
- We systematically evaluate 21 open-source, closed-source, and expert MLLMs. The experiments reveal the capability boundaries of MLLMs in surround-view driving video understanding and reasoning from both overall accuracy and fine-grained question-type perspectives.

2 Related Work

2.1 General Video Understanding Benchmarks

MVBench (Li et al., 2024b) converts public video annotations into multiple-choice QA tasks, covering 20 categories of video tasks that cannot be solved by relying on a single frame alone, thereby emphasizing temporal understanding from perception to cognition. Video-MME (Fu et al., 2025) further expands the coverage of general video evaluation by incorporating diverse open-domain video

Figure 2: Illustration of the MVDriveBench annotation pipeline.

categories and different video lengths with manually annotated QA pairs. LongVideoBench (Wu et al., 2024) and EgoSchema (Mangalam et al., 2023) respectively focus on long-context video understanding and long-horizon human activity understanding. In addition, TempCompass (Liu et al., 2024) is designed to evaluate temporal perception in videos covering temporal dimensions such as action, speed, direction, attribute change, and event order. These general video benchmarks are important for evaluating the video perception and temporal reasoning capabilities of MLLMs, but they are not specifically designed for autonomous driving scenarios.

2.2 Vision-Language Benchmarks and Datasets for Autonomous Driving

BDD-X (Kim et al., 2018) and LingoQA (Marcu et al., 2024) mainly rely on front-view videos, which limits their coverage of side-rear vehicles, rear-view risks, and blind spots. NuScenes-QA (Qian et al., 2024) constructs questions from structured annotations or scene graphs, but is closer to static scene understanding and lacks direct evaluation of synchronized video dynamics and cross-view object continuity. Talk2BEV (Choudhary et al., 2024) evaluates reasoning over BEV representations, but does not involve multi-view video reasoning.

Recent benchmarks, including DriveLM (Sima et al., 2024), Reason2Drive (Nie et al., 2024), and DriveLMM-o1 (Ishaq et al., 2025), emphasize multi-step reasoning, explanation, and planning. However, their questions do not necessarily force models to ground answers in multiple raw camera views, making it possible to rely on single-frame cues, structured representations, or language priors. DriveBench (Xie et al., 2025) focuses on visual grounding and reliability, but not on spatial-topological relationships among surround-view cameras. NuPlanQA (Park et al., 2025) and STSBench (Fruhwirth-Reisinger et al., 2026) are closer to multi-view driving understanding, yet they mainly focus on multi-view images, BEV-based scene understanding, or closed-form multiple-choice QA, leaving diverse surround-view video reasoning abilities insufficiently examined.

As summarized in Table 1, existing benchmarks still lack systematic evaluation of synchronized surround-view video reasoning. MVDriveBench addresses this gap by constructing QA tasks from synchronized six-view driving videos, requiring models to associate traffic participants across cameras, understand spatial continuity between adjacent views, infer temporal motion changes, and convert multi-view video evidence into spatio-temporal reasoning for safe driving.

3 MVDriveBench

As shown in Figure 2, the overall pipeline consists of three stages: surround-view-demand video curation, surround-view video rendering and QA

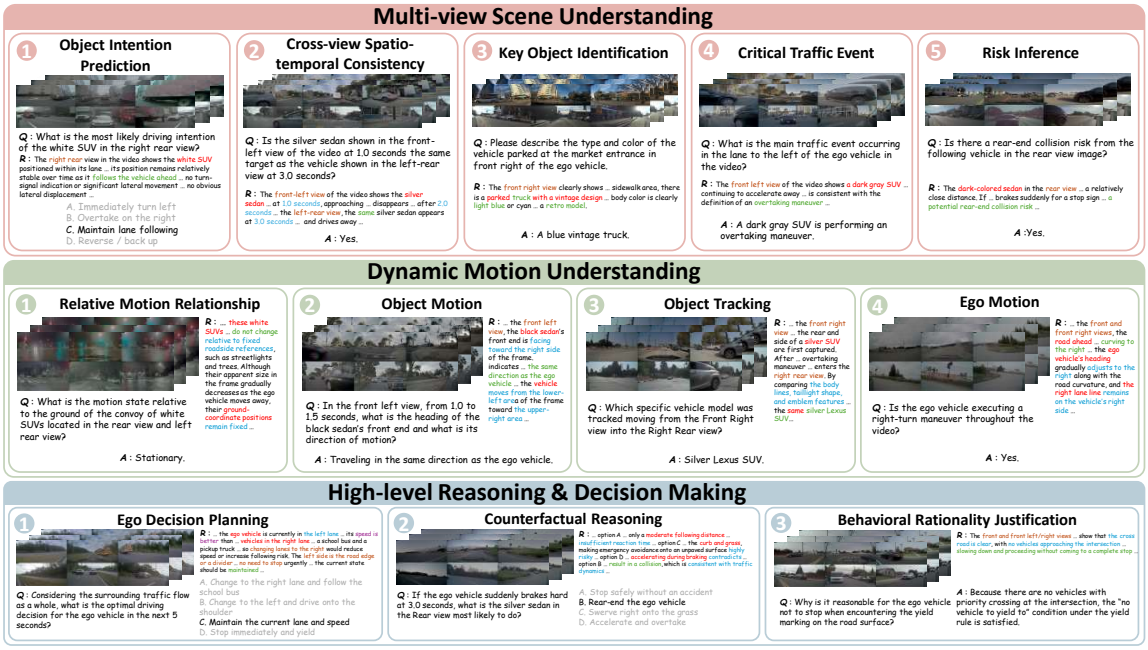


Figure 3: Illustration of the MVDriveBench's Evaluation Task.

generation, and manual quality filtering and data organization. Given a large-scale real-world autonomous driving video corpus $\mathcal{D} = \{c_i\}_{i=1}^{N}$, each clip c_i contains synchronized multi-camera observations $\mathbf{V}_i = \{V_i^k\}_{k=1}^{K}$, where $K = 6$ denotes the six surround-view cameras. We denote the front wide-angle camera stream as $F_i \in \mathbf{V}_i$. The goal of MVDriveBench is to construct a benchmark $\mathcal{B}$ consisting of high-quality question-answer pairs that cannot be reliably solved from F_i alone, but require surround-view spatial-temporal reasoning over $\mathbf{V}_i$.

3.1 Surround-View-Demand Video Curation

The video clips in MVDriveBench are sourced from the NVIDIA PhysicalAI-Autonomous-Vehicles dataset (NVIDIA, 2025). This dataset is a large-scale real-world autonomous driving dataset, covering diverse traffic flows, weather conditions, road environments, obstacles, pedestrians, and other driving-related elements.

The goal of the first stage is to select driving clips that truly require surround-view understanding. Unlike random sampling or directly using all video clips, we first determine whether each candidate clip provides sufficient spatio-temporal information for understanding the current driving state using only the front-view camera. Formally, for each clip c_i, only the front-view stream F_i is provided to a vision-language model with a structured filtering prompt. The model predicts

$$g(F_i) = (m_i, a_i, s_i, d_i, r_i), \tag{1}$$

where $m_i \in \{0, 1\}$ indicates whether surround-view information is necessary, $a_i \in \mathcal{A}$ denotes the driving behavior type of the ego vehicle, (s_i, d_i) represents the key temporal window, and r_i is the filtering rationale.

Specifically, we use Qwen3.5-397B-A17B (Qwen Team, 2026b) to analyze each front-view clip. The filtering prompt guides the model to reason step by step following the order of ego-vehicle motion state, surrounding traffic environment, necessity of surround-view information, and key temporal window localization. If the model determines that relying solely on the front-view perspective is insufficient to obtain the spatio-temporal information required for driving, such as observing side-rear blind spots during lane changes, assessing the motion states of vehicles in adjacent lanes during merging, or perceiving lateral oncoming traffic at intersections, the corresponding video clip is labeled as requiring surround-view understanding. The selected set is defined as

$$\mathcal{S} = \{c_i \in \mathcal{D} \mid m_i = 1\}. \tag{2}$$

This stage also supplements each selected clip with structured metadata, including clip ID, action type, filtering reason, key start time, and duration. The

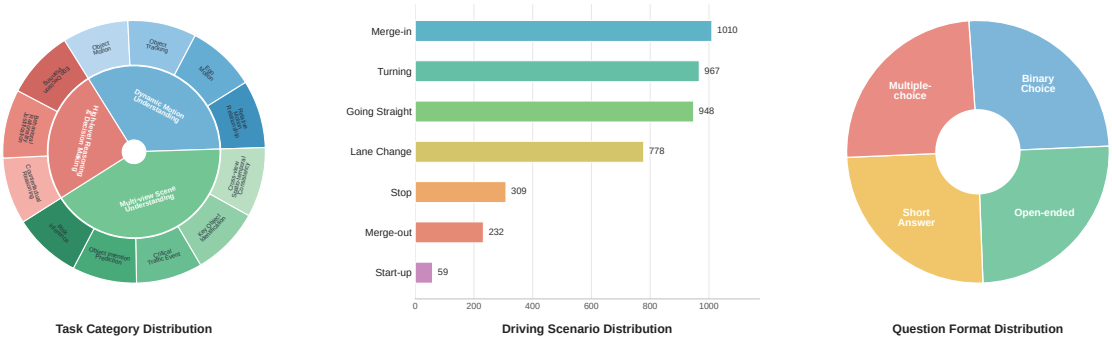


Figure 4: Statistical analysis of our MVDriveBench dataset. It consists of 3 high-level dimensions and 12 specific tasks.

output serves as the input for subsequent surround-view QA construction.

3.2 Surround-View Video Rendering

After obtaining candidate clips that require surround-view understanding, we construct a unified surround-view video input. For each selected clip $c_i \in \mathcal{S}$, we extract a fixed-length segment from the localized temporal window (s_i, d_i) and synchronously read six onboard camera views: Front Left, Front, Front Right, Left Rear, Rear, and Right Rear. These views are rendered into a unified 2×3 30 Hz grid video:

$$\tilde{\mathbf{V}}_i = R\left(\{V_i^k[s_i : s_i + d_i]\}_{k=1}^{6}\right), \qquad (3)$$

where $R(\cdot)$ denotes the surround-view rendering function. This representation allows the model to observe the front, side, and rear regions around the ego vehicle within a single video input.

The six-grid representation has two advantages. First, it preserves the temporal synchronization among different cameras, enabling the model to observe the continuous appearance, disappearance, and relative motion of objects across multiple views. Second, it provides a unified input format for cross-view spatial reasoning. For example, the model can judge the boundary continuity between the Left Rear and Rear views, or combine the Front Right and Right Rear views to determine whether the right-side blind spot is occupied.

3.3 Two-Stage Prompting for Surround-View QA Generation

Given the rendered surround-view video $\tilde{\mathbf{V}}_i$, we adopt a two-round prompting mechanism to generate high-quality QA pairs. Instead of directly asking the model to generate questions from the video, this design first requires the model to perform fine-grained scene understanding, and then generate structured QA pairs based on the intermediate understanding. This is important because surround-view driving videos contain complex spatial topology, object motion, and driving risk relationships; direct QA generation can easily lead to shallow questions or answers without sufficient visual evidence.

The first-round prompt focuses on scene understanding. The model takes the 5-second six-grid surround-view video as input and outputs a detailed scene analysis:

$$z_i = h_1(\tilde{\mathbf{V}}_i, p_1), \qquad (4)$$

where p_1 denotes the scene-understanding prompt, and z_i contains intermediate descriptions such as object locations, object motions, relative motion relationships, cross-view associations, temporal changes, bounding-box-level evidence, and potential risks. This prompt explicitly introduces the spatial topology rules of the six views, imaging change, relative motion analysis methods and temporal analysis principle, providing an intermediate semantic foundation for subsequent QA construction.

The second-round prompt focuses on VQA generation. Within the same dialog session, the model continues to generate 12 challenging QA pairs based on the scene understanding result from the first round:

$$\mathcal{Q}_i = h_2(z_i, p_2) = \{(q_{ij}, y_{ij}, \rho_{ij}, \tau_{ij}, \phi_{ij})\}_{j=1}^{12}, \qquad (5)$$

where p_2 denotes the QA-generation prompt, q_{ij} and y_{ij} represent the question and answer, ρ_{ij} represents the reasoning process of the question, $\tau_{ij} \in \mathcal{T}$ denotes the question type, $\phi_{ij} \in \mathcal{F}$ denotes the question format.

The type set $\mathcal{T}$ contains 12 fine-grained driving reasoning categories in Figure 3, covering the full reasoning chain from low-level visual perception to high-level driving decision-making. ***Relative Motion Relationships***, ***Object Motion***, ***Object Tracking*** and ***Ego-vehicle Motion*** focus on dynamic changes in videos. ***Intention Prediction***, ***Cross-view Spatio-Temporal Consistency***, ***Key Object Identification***, ***Critical Traffic Events***, and ***Risk Inference*** emphasize spatial topology and object continuity across multiple camera views. ***Ego-vehicle Decision Planning***, ***Counterfactual Reasoning*** and ***Behavior Rationality Justification*** further require the model to transform multi-view video evidence into safety-critical driving judgments.

To prevent the final benchmark from being dominated by a few common driving behaviors, we introduce action-type-based quota control during QA generation. Let n_a denote the number of generated clips for action type $a \in \mathcal{A}$, and let b_a denote its target quota. We only continue sampling clips of type a when

$$n_a < b_a. \tag{6}$$

As shown in Figure 4, this strategy encourages a balanced distribution of different ego-vehicle behaviors.

3.4 Manual Quality Check and Noise Filtering

Since the QA pairs are generated with the assistance of a strong model, they still contain issues such as object recognition errors, incorrect cross-view associations, non-unique answers, or insufficient reasoning evidence. Therefore, after the generation stage, we introduce manual quality checking to remove noisy questions. Let $\mathcal{E}_i$ denote the set of manually marked invalid question indices for clip c_i. The final benchmark is obtained by filtering the generated QA pairs:

$$\mathcal{B} = \bigcup_{c_i \in \mathcal{S}} \{(q_{ij}, y_{ij}, \rho_{ij}, \tau_{ij}, \phi_{ij}, a_i, c_i) \mid j \notin \mathcal{E}_i\}. \tag{7}$$

3.5 Evaluation Protocol

Given the constructed benchmark $\mathcal{B}$, each evaluated model M produces an inference result file $\mathcal{R}_M = \{\hat{y}_n\}_{n=1}^{|\mathcal{B}|}$, where $\hat{y}_n$ denotes the model response to the n-th question. For each valid sample, the evaluation function assigns a score according to its question format:

$$s_n = \mathrm{Eval}(q_n, y_n, \hat{y}_n, \rho_{ij}, \tau_n, \phi_n), \tag{8}$$

where y_n is the reference answer.

For closed-ended binary questions, we assign a binary score by exact matching:

$$s_n = \mathbb{I}\left[\mathrm{Norm}(\hat{y}_n) = \mathrm{Norm}(y_n)\right]. \tag{9}$$

For closed-ended multiple-choice questions, we extract the predicted option letter from the model response and compare it with the ground truth option, also yielding a $0/1$ score. For closed-ended short-answer and open-ended questions, we adopt an LLM-as-Judge protocol. ***Reliable Judge Model Qwen3-8B*** in C.4 (Bai et al., 2025a) receives question, reference answer, model answer and the reasoning for the question, and outputs an evaluation result containing a scalar score and a brief reason:

$$(s_n, r_n) = J(q_n, y_n, \hat{y}_n, \rho_{ij}, \tau_n, \phi_n), \quad s_n \in [0, 1]. \tag{10}$$

The overall accuracy of model M is computed as the mean score over all valid samples:

$$\mathrm{Acc}(M) = \frac{1}{|\mathcal{R}_M^{\mathrm{valid}}|} \sum_{n \in \mathcal{R}_M^{\mathrm{valid}}} s_n. \tag{11}$$

4 Main Results

Table 2 shows the performance of 21 MLLMs on MVDriveBench. From the model perspective, closed-source models still maintain an overall lead, yet a clear gap remains between current MLLMs and human performance. GPT-5 (Singh et al., 2026) achieves the highest *Overall* score of 78.7, but still trails human performance (88.5) by 9.8 points, indicating that spatio-temporal understanding and reasoning over surround-view driving videos remain highly challenging for existing MLLMs. In contrast, Specialist Models do not exhibit the expected advantages on MVDriveBench. Driving- or embodied-oriented models such as ReCogDrive (Li et al., 2026), Cosmos (NVIDIA Corporation, 2026a), and MiMo (Xiaomi Corporation, 2026) obtain *Overall* scores only in the range of 45.7–54.4, which are substantially lower than those of closed-source models and also below most large-scale open-source models. A closer inspection shows that these models generally achieve high scores on ego-vehicle decision planning tasks. For example, ReCogDrive-VLM-8B, Cosmos-Reason1-7B, and MiMoEmbodied-7B reach 97.8, 98.6, and 95.3, respectively. However, their average scores on Dynamic Motion Understanding are only 28.1–34.0, and their average scores on Multi-view Scene Understanding

Model	Overall	Dynamic Motion Understanding					Multi-view Scene Understanding						Driving Reasoning			
		Rel. Mot.	Obj. Mot.	Tracking	Ego Mot.	Avg	Intent	Cross-view	Key Obj.	Key Event	Risk	Avg	Planning	Counterfactual	Rationality	Avg
Human	88.5	87.5	80.3	90.1	88.8	86.7	77.7	94.1	90.1	81.6	92.3	87.2	95.9	94.1	89.5	93.2
Closed Source Models																
GPT-5 (Singh et al., 2026)	**78.7**	58.9	57.5	62.7	87.2	66.6	83.0	75.7	72.3	74.2	91.5	79.3	93.9	93.4	**93.5**	93.6
GPT-4o (OpenAI et al., 2024)	74.3	48.9	46.3	50.7	70.4	54.1	**85.3**	89.4	67.9	70.2	87.8	80.1	97.0	92.9	84.6	91.5
Gemini-2.5-Pro (Comanici et al., 2025)	76.7	60.8	56.4	57.1	87.7	65.5	68.6	70.5	74.2	71.6	92.2	75.4	95.2	**96.7**	88.2	93.4
Gemini-3-Pro (Google DeepMind, 2025)	78.3	61.3	52.9	64.3	86.1	66.1	83.4	78.0	74.0	69.8	**95.1**	80.1	93.9	94.8	85.6	91.4
Open Source MoE Models																
Qwen3.5-397B-A17B (Qwen Team, 2026b)	74.7	57.9	60.0	54.1	**89.0**	65.2	78.9	70.0	69.0	66.5	88.6	74.6	96.1	86.7	79.8	87.5
ERNIE-4.5-VL-28B-A3B-Thinking (Baidu-ERNIE-Team, 2025)	49.9	24.5	27.4	34.1	52.1	34.5	53.8	17.8	48.9	41.2	80.9	48.5	74.6	75.1	69.5	73.1
Qwen3-VL-30B-A3B-Thinking (Bai et al., 2025a)	66.4	49.0	58.9	57.7	52.1	54.4	80.7	14.6	70.6	70.1	81.4	63.5	92.0	90.1	81.6	87.9
Open Source Models > 7B																
InternVL3-78B (Zhu et al., 2025)	71.2	48.7	60.0	49.2	78.2	59.0	66.7	**91.1**	62.5	62.3	78.9	72.3	97.2	91.5	66.9	85.2
Qwen2.5-VL-72B (Bai et al., 2025b)	69.4	43.2	49.1	42.3	74.9	52.4	79.8	63.5	61.3	62.6	88.9	71.2	97.8	90.4	79.1	89.1
InternVL3-38B (Zhu et al., 2025)	70.9	53.2	56.0	50.3	79.1	59.6	63.7	86.2	59.7	61.1	85.6	71.3	98.1	90.7	66.1	85.0
Qwen2.5-VL-32B (Bai et al., 2025b)	71.5	49.9	48.3	37.6	79.1	53.7	76.0	83.8	63.8	57.3	91.1	74.4	98.3	90.4	81.2	90.0
Qwen3.5-27B (Qwen Team, 2026a)	61.7	**63.2**	61.4	59.3	52.1	59.0	**26.3**	14.6	**81.6**	**80.7**	81.2	56.9	55.5	73.7	89.9	73.0
GLM-4.1V-9B-Thinking (Team et al., 2026)	**43.2**	28.1	34.6	37.4	54.3	38.6	32.5	20.5	54.3	33.8	81.4	44.5	**29.0**	**47.0**	64.2	46.7
Video-R1-7B (Feng et al., 2026)	62.2	55.2	**73.4**	**73.4**	69.2	67.8	51.5	15.1	24.2	53.0	74.2	43.6	93.9	80.5	73.4	82.6
LLaVA-OneVision-Qwen2-7B-OV (Li et al., 2024a)	50.5	26.2	22.9	33.8	76.0	39.7	47.1	35.1	42.1	35.8	86.1	49.2	92.0	80.5	28.1	66.9
VGLM-QA-VGGT-8B (Zheng et al., 2026)	56.4	43.7	50.3	38.2	71.9	51.0	62.0	27.6	38.3	38.7	86.4	50.6	94.8	77.3	48.2	73.4
Specialist Models																
ReCogDrive-VLM-8B (Li et al., 2026)	47.3	**14.5**	14.9	**12.1**	73.8	28.8	69.0	61.1	18.0	16.6	86.7	50.3	97.8	78.8	24.2	66.9
Cosmos-Reason1-7B (NVIDIA et al., 2025)	45.7	15.6	**13.1**	15.9	67.8	28.1	69.6	82.7	**12.2**	**16.6**	**51.0**	46.4	**98.6**	84.1	**20.6**	67.8
Cosmos-Reason2-8B (NVIDIA Corporation, 2026b)	51.3	16.7	14.3	21.4	83.7	34.0	69.9	77.3	23.2	20.5	76.5	53.5	95.9	91.5	23.8	70.4
MiMoEmbodied-7B (Hao et al., 2026)	47.3	24.5	22.6	20.1	**52.1**	29.8	79.8	**14.6**	30.0	26.0	81.2	46.3	95.3	87.3	36.7	73.1
MiMo-VL-7B-SFT (Team et al., 2025)	54.4	18.1	15.2	20.1	76.3	32.4	82.7	83.5	27.1	24.6	88.4	61.3	94.5	86.1	35.7	72.1

Table 2: Surround-view reasoning results on MVDriveBench. **Bold** indicates the best performance, while **underline** indicates the worst performance.

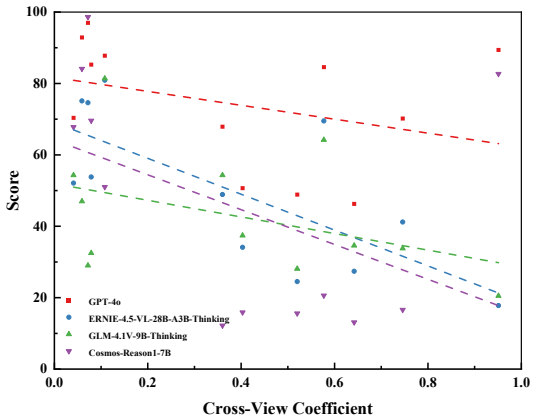


Figure 5: The trend between the proportion of multi-view questions and the response scores.

Figure 6: Heatmap of one bad case. Query: Summarize the surrounding traffic flow based on the surround-view video. Answer: Less traffic flow. GT: Heavy traffic due to a large number of vehicles.

are merely 46.3–61.3. This suggests that their strong planning performance is not grounded in reliable surround-view perception or cross-view spatio-temporal reasoning.

From the task perspective, Driving Reasoning tasks are generally the easiest for models to score highly on, but this does not imply that current models have acquired a comprehensive driving understanding. Several Specialist Models exceed 95 points on planning tasks, while their *Overall* scores remain below 55, indicating that they rely more on generic driving priors rather than completing a full perception–reasoning–decision chain based on evidence from surround-view videos. By contrast, Dynamic Motion Understanding is substantially more challenging: even the best-performing closed-source model achieves an average dynamic understanding score of only 66.6, still far below the human score of 86.7. In Multi-view Scene Understanding, the gap in cross-view spatio-temporal consistency is particularly pronounced. For instance, GPT-4o (OpenAI et al., 2024) and InternVL3-78B (Zhu et al., 2025) achieve 89.4 and 91.1, respectively, whereas Qwen3-VL-30B-A3B-Thinking (Bai et al., 2025a), Qwen3.5-27B (Qwen Team, 2026a), and MiMoEmbodied-7B (Hao et al., 2026) obtain only 14.6, resulting in a maximum gap of 76.5 points. Figure 6 visualizes that the position of Qwen3.5-27B's attention does not fall on surrounding traffic vehicles. These results indicate that cross-camera object association and multi-view spatio-temporal consistency modeling remain core bottlenecks in current surround-view driving video understanding.

As shown in Figure 5, we compute the proportion of questions in each of the 12 task categories that require multi-view information for reliable answering, and define it as the cross-view coefficient. Through linear fitting, we observe a negative correlation between the cross-view coefficient and model performance. This indicates that current MLLMs tend to perform worse on tasks that rely

Model	Overall	Dynamic Motion Understanding					Multi-view Scene Understanding						Driving Reasoning			
		Rel. Mot.	Obj. Mot.	Tracking	Ego Mot.	Avg	Intent	Cross-view	Key Obj.	Key Event	Risk	Avg	Planning	Counterfactual	Rationality	Avg
Closed Source Models																
GPT-5	77.2↓ 1.5	64.9	63.5	54.4	87.9	67.7	75.0	63.5	61.7	78.1	96.6	75.0	96.5	90.7	94.5	93.9
GPT-4o	65.1↓ 9.2	41.5	35.7	25.0	77.7	45.0	77.2	42.7	54.6	68.6	85.0	65.6	98.1	90.7	85.3	91.4
Gemini-2.5-Pro	73.2↓ 3.5	59.4	49.1	55.2	84.8	62.1	66.0	67.1	61.9	72.0	87.7	70.9	96.9	89.8	87.7	91.5
Gemini-3-Pro	78.4↑ 0.1	50.8	55.2	65.2	89.6	65.2	79.2	87.3	59.2	75.6	92.3	78.7	100.0	94.8	87.8	94.2
Open Source MoE Models																
Qwen3.5-397B-A17B	59.6↓ 15.1	40.1	46.6	29.4	75.2	47.8	73.4	37.0	33.4	46.6	89.2	55.9	90.6	76.8	77.5	81.6
ERNIE-4.5-VL-28B-A3B-Thinking	43.6↓ 6.3	21.4	26.3	16.5	52.1	29.1	43.3	14.9	33.6	34.4	80.9	41.4	70.4	66.3	63.8	66.8
Qwen3-VL-30B-A3B-Thinking	53.6↓ 12.8	33.1	31.4	24.5	52.1	35.3	71.6	14.3	35.1	52.2	81.2	50.9	92.3	81.3	75.5	83.0
Open Source Models >7B																
InternVL3-78B	55.6↓ 15.6	26.7	36.0	20.6	67.5	37.7	66.7	40.5	32.0	50.6	82.3	54.4	97.0	84.7	63.0	81.6
Qwen2.5-VL-72B	52.4↓ 17.0	20.3	30.3	15.1	63.9	32.4	82.2	16.5	23.2	44.4	84.8	50.2	98.3	84.1	68.1	83.5
InternVL3-38B	55.9↓ 15.0	34.5	36.0	21.4	68.0	40.0	71.3	25.9	35.0	48.2	85.3	53.1	98.6	89.0	59.1	82.2
Qwen2.5-VL-32B	56.2↓ 15.3	34.5	35.1	19.2	70.5	39.8	74.9	21.9	29.5	49.9	85.9	52.4	96.7	83.9	74.5	85.0
Qwen3.5-27B	47.5↓ 14.2	48.5	55.7	39.6	52.1	49.0	15.8	14.3	47.7	59.7	81.2	43.7	25.1	44.8	85.0	51.6
GLM-4.1V-9B-Thinking	47.2↑ 4.0	17.3	21.4	13.7	52.1	26.1	54.4	14.3	36.9	35.9	81.2	44.5	85.9	76.2	77.4	79.8
Video-R1-7B	51.1↓ 11.1	33.7	54.9	58.0	66.7	53.3	51.2	15.4	26.0	51.2	84.5	45.7	91.7	78.5	76.2	82.1
LLaVA-OneVision-Qwen2-7B-OV	47.1↓ 3.4	27.9	26.3	22.3	66.9	35.9	48.0	23.0	30.3	34.8	84.8	44.2	90.9	79.3	31.7	67.3
VGLM-QA-VGGT-8B	51.3↓ 5.1	36.2	31.1	40.1	67.8	43.8	63.7	20.8	30.8	25.6	83.7	44.9	93.9	76.5	45.7	72.0
Specialist Models																
ReCogDrive-VLM-8B	44.4↓ 2.9	15.0	14.9	12.1	67.2	27.3	70.5	45.1	14.7	17.0	88.9	47.2	94.2	70.0	23.9	62.7
Cosmos-Reason1-7B	44.1↓ 1.6	10.0	6.3	12.1	74.1	25.6	72.8	58.6	11.1	13.2	82.0	47.5	97.0	77.1	15.7	63.3
Cosmos-Reason2-8B	42.9↓ 8.4	12.8	9.4	9.6	70.0	25.4	75.7	28.9	10.6	15.0	86.7	43.4	94.2	82.7	20.7	65.9
MiMoEmbodied-7B	42.8↓ 4.5	19.5	14.3	12.4	52.3	24.6	74.3	14.3	18.2	19.4	80.6	41.4	95.3	84.7	30.6	70.2
MiMo-VL-7B-SFT	44.9↓ 9.5	13.6	10.6	8.2	71.9	26.1	78.4	39.5	14.6	17.8	85.0	47.1	95.0	80.2	25.5	66.9

Table 3: Front-view ablation reasoning results on MVDriveBench.

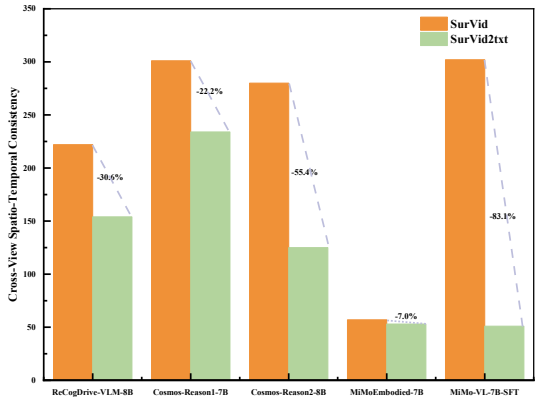


Figure 7: Performance comparison before and after converting surround-view videos into caption descriptions.

more heavily on multi-view perception and reasoning.

5 Ablation Study

Front-View Ablation: As shown in Table 3, when only front-view inputs are provided, most models exhibit a substantial drop in *Overall* performance. In particular, Qwen2.5-VL-72B (Bai et al., 2025b), InternVL3-78B (Zhu et al., 2025), Qwen2.5-VL-32B (Bai et al., 2025b), and Qwen3.5-397B-A17B (Qwen Team, 2026b) decrease by 17.0, 15.6, 15.3, and 15.1 points, respectively, indicating that the front-view perspective alone is insufficient to capture the complete surrounding traffic state.

At the task level, Dynamic Motion Understanding and Multi-view Scene Understanding are the most affected categories. Tasks such as object tracking, key object identification, and cross-view spatio-temporal consistency are highly dependent on information from side and rear cameras. For example, the Cross-view score of InternVL3-78B (Zhu et al., 2025) drops from 91.1 to 40.5, while the Key Object score of Qwen2.5-VL-72B (Bai et al., 2025b) decreases from 61.3 to 23.2. These results suggest that surround-view visual evidence is crucial for robust spatio-temporal understanding and reasoning.

Meanwhile, Driving Reasoning tasks still maintain relatively high scores under the front-view-only setting, with several models achieving planning scores above 95. This indicates that models generate plausible answers by relying on generic driving priors rather than genuinely understanding the complete surround-view scene. Overall, this ablation study shows that the question design of MVDriveBench indeed exhibits strong multi-view dependency and can effectively distinguish whether a model reasons based on real surround-view spatio-temporal evidence or merely guesses from front-view information and language priors.

SurVid2txt Ablation: We convert the surround-view videos into caption descriptions and require MLLMs to answer the evaluation questions using only these textual descriptions. As shown in Figure 7, all models exhibit varying degrees of performance degradation on tasks with the highest cross-view coefficients.

6 Conclusion

We propose **MVDriveBench** to evaluate MLLMs' surround-view video reasoning in autonomous driving. Extensive experiments show that existing MLLMs still struggle with robust multi-view spatio-temporal understanding.

Limitations

Our QA construction pipeline relies on strong MLLMs for clip filtering, scene understanding, and initial QA generation. Although we introduce manual quality checking to remove unsupported or erroneous samples, model-assisted generation may still introduce subtle biases in question style, answer distribution, reasoning patterns, or scenario selection. In particular, some generated questions may reflect the prior knowledge or failure modes of the generation model, rather than purely human-designed evaluation objectives. Future work may extend MVDriveBench toward human-centered evaluation protocols.

Ethical considerations

We checked the collected and used data for personally identifying information and offensive content during the data construction and manual verification process. Since MVDriveBench is built from autonomous-driving videos, the raw data may contain road users, vehicles, license plates, or other scene elements from public traffic environments. We do not use or annotate personal names, identities, contact information, or any attributes intended to uniquely identify individuals. During manual quality checking, annotators were instructed to filter out QA pairs containing offensive, discriminatory, abusive, or otherwise inappropriate content. The released benchmark focuses only on traffic participants, driving behaviors, scene dynamics, and safety-related reasoning, and does not include personally identifying descriptions of individuals.

We provide compensation to all annotators involved in the manual annotation and quality verification process. The compensation is calculated based on the completed and verified annotation workload, aiming to reflect the time and effort required to review surround-view videos, verify QA pairs, and correct annotation errors. Annotators receive approximately $0.5 for each sample they verify.

MVDriveBench is constructed based on the PhysicalAI-Autonomous-Vehicles dataset (NVIDIA, 2025), whose use is governed by the NVIDIA Autonomous Vehicle Dataset License Agreement. Accordingly, our use of the source videos follows the permitted autonomous-vehicle-related research purpose specified by the original dataset license. We do not redistribute the raw source videos beyond the terms allowed by the original data provider. Any released benchmark artifacts, such as annotations, question-answer pairs, metadata, prompts, and evaluation scripts, will be distributed with a clear license statement and will require users to comply with the license and usage restrictions of the underlying NVIDIA dataset. For third-party models, APIs, and toolkits used in our experiments, we follow their respective licenses, terms of service, and attribution requirements.

Generative AI tools were used solely for language polishing, grammar correction, and improving the readability of this manuscript. All research ideas, methods, experiments, analyses, and conclusions were conducted and verified by the authors, who take full responsibility for the final content.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Baidu-ERNIE-Team. 2025. Ernie 4.5 technical report. https://ernie.baidu.com/blog/publication/ERNIE_Technical_Report.pdf.

Tushar Choudhary, Vikrant Dewangan, Shivam Chandhok, Shubham Priyadarshan, Anushka Jain, Arun K Singh, Siddharth Srivastava, Krishna Murthy Jatavallabhula, and K Madhava Krishna. 2024. Talk2bev: Language-enhanced bird's-eye view maps for autonomous driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16345–16352. IEEE.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, Tianren Gao, Erlong Li, Kun Tang, Zhipeng Cao, Tong Zhou, Ao Liu, Xinrui Yan, Shuqi Mei, Jianguo Cao, and 2 others. 2024. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 958–979.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. 2026. Video-r1: Reinforcing video reasoning in mllms. *Advances in Neural Information Processing Systems*, 38:99114–99137.

Christian Fruhwirth-Reisinger, Dusan Malic, Wei Lin, David Schinagl, Samuel Schulter, and Horst Possegger. 2026. Stsbench: A spatio-temporal scenario benchmark for multi-modal large language models in autonomous driving. *Advances in Neural Information Processing Systems*, 38.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, and 2 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *Preprint*, arXiv:2405.21075.

Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, Caifeng Shan, and Ran He. 2024. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *Preprint*, arXiv:2411.15296.

Google DeepMind. 2025. Gemini 3 pro model card. `https://deepmind.google/models/model-cards/`. Updated: 18 November 2025. Accessed: 2026-05-21.

Xiaoshuai Hao, Lei Zhou, Zhijian Huang, Zhiwen Hou, Yingbo Tang, Lingfeng Zhang, Guang Li, Zheng Lu, Shuhuai Ren, Xianhui Meng, Yuchen Zhang, Jing Wu, Jinghui Lu, Chenxu Dang, Jiayi Guan, Jianhua Wu, Zhiyi Hou, Hanbing Li, Shumeng Xia, and 25 others. 2026. Mimo-embodied: X-embodied foundation model technical report. *Preprint*, arXiv:2511.16518.

Ayesha Ishaq, Jean Lahoud, Ketan More, Omkar Thawakar, Ritesh Thawkar, Dinura Dissanayake, Noor Ahsan, Yuhao Li, Fahad Shahbaz Khan, Hisham Cholakkal, Ivan Laptev, Rao Muhammad Anwer, and Salman H. Khan. 2025. Drivelmm-o1: A step-by-step reasoning dataset and large multimodal model for driving scenario understanding. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 20501–20508. IEEE.

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

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. *Preprint*, arXiv:2311.17005.

Yongkang Li, Kaixin Xiong, Xiangyu Guo, Fang Li, Sixu Yan, Gangwei Xu, Lijun Zhou, Long Chen, Haiyang Sun, BING WANG, Kun Ma, Guang Chen, Hangjun Ye, Wenyu Liu, and Xinggang Wang. 2026. Recogdrive: A reinforced cognitive framework for end-to-end autonomous driving. In *The Fourteenth International Conference on Learning Representations*.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8731–8772.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Preprint*, arXiv:2308.09126.

Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoît Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, Elahe Arani, and Oleg Sinavski. 2024. Lingoqa: Visual question answering for autonomous driving. In *European Conference on Computer Vision*, pages 252–269. Springer.

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

NVIDIA, :, Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, Liang Feng, Francesco Ferroni, Rama Govindaraju, Jinwei Gu, Siddharth Gururani, Imad El

Hanafi, Zekun Hao, and 35 others. 2025. Cosmos-reason1: From physical common sense to embodied reasoning. *Preprint*, arXiv:2503.15558.

NVIDIA. 2025. Physicalai-autonomous-vehicles. https://huggingface.co/datasets/nvidia/PhysicalAI-Autonomous-Vehicles. Hugging Face dataset. Accessed: 2026-05-21.

NVIDIA Corporation. 2026a. About nvidia. https://www.nvidia.com/en-sg/about-nvidia/. Official company website. Accessed: 2026-05-21.

NVIDIA Corporation. 2026b. Cosmos-reason2-8b. https://huggingface.co/nvidia/Cosmos-Reason2-8B. Hugging Face model card. Accessed: 2026-05-21.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Sung-Yeon Park, Can Cui, Yunsheng Ma, Ahmadreza Moradipari, Rohit Gupta, Kyungtae Han, and Ziran Wang. 2025. Nuplanqa: A large-scale dataset and benchmark for multi-view driving scene understanding in multi-modal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8066–8076.

Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. 2024. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4542–4550.

Qwen Team. 2026a. Qwen3.5-27b. https://huggingface.co/Qwen/Qwen3.5-27B. Hugging Face model card. Accessed: 2026-05-21.

Qwen Team. 2026b. Qwen3.5-397b-a17b-fp8. https://huggingface.co/Qwen/Qwen3.5-397B-A17B-FP8. Hugging Face model card. Accessed: 2026-05-21.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. 2024. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pages 256–274. Springer.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, and 55 others. 2025. Mimo-vl technical report. *Preprint*, arXiv:2506.03569.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 74 others. 2026. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Preprint*, arXiv:2407.15754.

Xiaomi Corporation. 2026. Company profile. https://ir.mi.com/corporate-information/company-profile. Official company profile. Accessed: 2026-05-21.

Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. 2025. Are vlms ready for autonomous driving? an empirical study from the reliability, data and metric perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6585–6597.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. 2026. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. *Advances in neural information processing systems*, 38:20560–20586.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

Appendices

A Task Definition

MVDriveBench consists of **12** question-answering tasks organized into three major categories: Dynamic Motion Understanding, Multi-view Scene Understanding, and Driving Reasoning. These tasks are designed to evaluate whether MLLMs can

perceive object-level motion, integrate surround-view spatio-temporal information, and perform safety-critical driving reasoning. Table 5 shows examples of each task type.

Relative Motion Relationships (Rel. Mot.) asks models to judge the relative motion state between the ego vehicle and surrounding objects, such as whether they are approaching or moving away.

Object Motion (Obj. Mot.) focuses on the motion direction or state of a specific traffic participant within a given view.

Object Tracking (Tracking) requires models to continuously track the same object across time and possibly across different camera views.

Ego-vehicle Motion (Ego Mot.) evaluates whether models can infer the motion state of the ego vehicle from visual changes in the environment.

Intention Prediction (Intent) requires predicting the likely intention of surrounding vehicles or pedestrians based on their visual state and context.

Cross-view Spatio-Temporal Consistency (Cross-view) evaluates whether models can associate the same object across adjacent camera views over time.

Key Object Identification (Key Obj.) asks models to identify the most important object affecting the ego vehicle's behavior.

Critical Traffic Events (Key Event) focuses on recognizing key driving events occurring in the video.

Risk Inference (Risk) evaluates whether models can infer potential safety risks from the surrounding traffic environment.

Ego-vehicle Decision Planning (Planning) asks models to select the safest or most appropriate driving action for the ego vehicle.

Counterfactual Reasoning (Counterfactual) evaluates whether models can reason about possible outcomes under hypothetical alternative actions.

Behavior Rationality Justification (Rationality) requires models to explain why the ego vehicle's observed behavior is reasonable or safe under the given traffic context.

B Annotation Details

Clip Selection for Surround-View Necessity Assessment: To identify driving scenarios that inherently require multi-view (surround-view) perception for safe navigation, we employ a large vision-language model (VLM) as an automated filter over the raw front-view video clips. Specifically, each 20-second clip—sampled at 30 Hz from the front-wide camera (120° FOV)—is presented to the VLM along with a structured analysis prompt. The prompt in Figure 8 guides the model through a three-step reasoning process: (1) identifying the ego vehicle's current maneuver (e.g., lane change, intersection traversal, merging); (2) characterizing the surrounding traffic environment (e.g., adjacent vehicles, pedestrians, traffic density); and (3) determining whether front-view information alone suffices for safe driving or whether critical blind-spot information necessitates surround-view perception. For clips classified as requiring multi-view capability, the model further predicts the temporal window (start second and duration) within the clip where the need is most acute, as well as the dominant action type. The output is structured as a JSON object with fields including need_multiview (boolean), reason, start_sec, duration_sec, and action_type. We adopt a safety-conservative design: when any uncertainty or potential risk exists, the model is instructed to favor a "needs multi-view" judgment. Clips with parsing failures or API exceptions are flagged separately and excluded from downstream evaluation. This filtering stage significantly reduces the annotation workload while ensuring that the resulting benchmark focuses on scenarios where multi-view perception provides genuine and measurable safety value.

Multi-View VQA Construction: The question-answer (QA) pairs for each selected clip are generated through a two-turn VLM prompting procedure. In the first turn, the VLM receives a 5-second, 30 Hz six-view concatenated surround-view video (rendered as a 2×3 grid comprising Front Left, Front, Front Right, Left Rear, Rear, and Right Rear) together with a detailed system prompt in Figure 9 that establishes the expert role, the six-view spatial topology rules (including the telephoto nature of the Rear camera and the topological continuity between adjacent views), the imaging principles for inferring relative motion from visual cues (size change, appearance/disappearance, vehicle orientation, and cross-view consistency), bounding-box analysis guidelines, and temporal segmentation rules requiring all state, motion, behavior, and relational conclusions to be timestamped. In the second turn, the VLM receives a task prompt in Figure 10 that instructs it to generate exactly 12 high-difficulty VQA pairs covering a fixed set of

question types: cross-view spatiotemporal consistency (closed-ended binary), target motion (closed-ended short-answer), target tracking (closed-ended short-answer), ego-vehicle motion (closed-ended binary), relative motion relationship (closed-ended short-answer), target intent prediction (closed-ended multiple-choice), important target identification (open-ended), important traffic event (open-ended), counterfactual reasoning (closed-ended multiple-choice), behavioral rationality argumentation (open-ended), risk inference (closed-ended binary), and ego-vehicle decision planning (closed-ended multiple-choice). The prompt enforces a balanced distribution of 3 binary, 3 short-answer, 3 multiple-choice, and 3 open-ended questions. Critically, the prompt requires that each QA pair be supported by an observable video evidence chain combined with at least one piece of external driving/traffic/geometry/kinematics knowledge, ensuring that questions cannot be answered by common sense alone. An upstream scene understanding result is provided as a hidden internal reference for cross-checking, but the final output must be "source-traceless"—all questions, answers, and reasoning must appear as if derived solely from the video. At least one of the 12 QA pairs must explicitly reference bounding-box coordinate information. For multiple-choice questions, the prompt mandates that distractors be semantically co-level with the correct option, include at least one plausible-but-disprovable choice, and that the reasoning process either sequentially eliminates options or applies counterfactual reasoning to each candidate. Each QA pair includes a reasoning field of no fewer than 4 sentences with at least 3 explicit reasoning steps, using video-source expressions throughout. The pipeline supports two video rendering modes—clean (without bounding-box overlays) and box (with bounding-box annotations)—and processes clips in parallel with per-action-type quotas (e.g., straight driving, lane change, turning, merging, exiting, stopping, starting, reversing, U-turn) to ensure balanced scenario coverage in the final benchmark.

Manual Annotation: To ensure the quality and reliability of the annotations, we recruited ten volunteers with diverse educational backgrounds and varying levels of expertise in autonomous driving. The annotators include participants with different professional experiences, ranging from junior researchers to domain experts, providing broad coverage of knowledge levels and annotation perspectives. Before annotation, all volunteers were provided with detailed annotation guidelines and trained to use our annotation tool, as illustrated in Figure 11. This procedure helps standardize the annotation process and reduce potential inconsistencies across annotators.

C Evaluation Details

MVDriveBench is constructed from the NVIDIA PhysicalAI-Autonomous-Vehicles dataset (NVIDIA, 2025), which provides synchronized multi-camera driving videos collected in real-world autonomous driving scenarios. Each video clip contains six onboard camera views, including front, rear, front-left, front-right, rear-left, and rear-right views. To provide a unified surround-view input for multimodal large language models (MLLMs), we concatenate the six synchronized camera streams into a 2×3 grid video. This format enables models to jointly observe the traffic environment in front of, beside, and behind the ego vehicle, which is essential for evaluating surround-view spatio-temporal understanding.

Based on these surround-view videos, we construct 4,303 question-answer pairs covering 12 task categories. These tasks span a complete driving reasoning pipeline, including cross-view spatio-temporal consistency, object tracking, ego-vehicle decision planning, counterfactual reasoning, and other perception- and reasoning-oriented capabilities. Each question is associated with one of four answer formats: binary Yes/No, short answer, multiple choice, and open-ended free-form answer. This design allows MVDriveBench to evaluate both automatically verifiable recognition tasks and more complex reasoning tasks that require semantic judgment.

C.1 Inference Prompt Design

During inference, each surround-view grid video is combined with a unified prompt template. The overall prompt structure is illustrated in Figure 12. Before injecting the specific `{question_format}` and `{question}`, the template first explicitly defines the output requirements for all four answer formats. For binary questions, the model is instructed to output only `Yes` or `No`; for multiple-choice questions, it must output only one option letter from `A/B/C/D`; for short-answer and open-ended questions, the model is also constrained to provide only

the final answer itself.

This prompt design serves two purposes. First, it standardizes the model outputs across heterogeneous MLLMs, reducing unnecessary explanations or format deviations. Second, it enables automatic parsing and evaluation for closed-form questions, especially binary and multiple-choice tasks. By enforcing the instruction that the model should "only output the answer itself without additional explanation," MVDriveBench minimizes evaluation noise caused by verbose responses rather than actual reasoning errors.

C.2 Ablation Settings

In addition to the main evaluation using full six-view surround videos, we design two ablation settings to analyze the necessity of visual coverage and the role of video input.

Ablation 1: Front-view-only input. In the first ablation, the model is given only the front-view video instead of the full surround-view grid. We use a separate prompt template in Figure 13. Since many questions in MVDriveBench require side-view or rear-view information, the template explicitly informs the model that these views are missing and allows it to make reasonable guesses when necessary. This setting evaluates whether front-view information alone is sufficient for solving surround-view driving reasoning tasks.

Ablation 2: Text-only QA based on generated captions. The second ablation removes direct video input during question answering. Specifically, we first ask the model to generate a textual description of the video. Then, the generated caption is used as the only context for answering the question. The corresponding templates are shown in Figure 14 and Figure 7. This setting evaluates the extent to which MLLMs can solve MVDriveBench through text-based reasoning alone after compressing visual evidence into a caption.

C.3 Evaluation Protocol

As shown in Figure 16 and 17, MVDriveBench adopts different evaluation strategies according to the answer format. For binary Yes/No questions and multiple-choice questions, we use exact matching. The prediction is assigned a score of 1 if it exactly matches the ground-truth answer, and 0 otherwise. This strict matching protocol is reliable because the inference prompt constrains the model to output only standardized labels.

Question Type	#Samples	Agreement / Pearson	Cohen's κ / Spearman	MAE
Short Answer	500	93.2%	0.86	–
Open-ended	500	0.88	0.86	0.061

Table 4: Agreement between human evaluation and LLM-as-Judge evaluation.

For short-answer and open-ended questions, exact string matching is insufficient because multiple semantically equivalent answers may be valid. Therefore, we adopt an LLM-as-Judge evaluation strategy using Qwen3-8B (Bai et al., 2025a) as the judge model. For short-answer questions, the judge determines whether the model prediction is semantically equivalent to the reference answer, producing a binary score of 0 or 1. For open-ended questions, the judge assigns a fine-grained score from 0.0 to 1.0 according to the semantic correctness, completeness, and consistency of the predicted answer with the reference. This hybrid evaluation protocol combines the reliability of exact matching for closed-form questions with the flexibility of semantic evaluation for free-form reasoning questions.

C.4 Validation of LLM-as-Judge Evaluation

Since short-answer and open-ended questions cannot be reliably evaluated by exact string matching, we further conduct a quantitative validation to examine the reliability of the LLM-as-Judge protocol. Specifically, we randomly sample a subset of model responses from short-answer and open-ended questions and ask human annotators to independently evaluate their correctness according to the same evaluation criteria. We then compare the scores assigned by human annotators with those produced by the judge model Qwen3-8B.

For short-answer questions, both human annotators and the judge model provide binary correctness labels, and we measure their agreement using accuracy and Cohen's kappa. For open-ended questions, where the score is continuous in the range of $[0, 1]$, we compute Pearson correlation and Spearman rank correlation between human scores and LLM-judge scores. The results are reported in Table 4. The high agreement and correlation indicate that the LLM-as-Judge protocol provides a stable and scalable approximation of human evaluation for semantic-answer questions in MVDriveBench.

C.5 Evaluated Models

We evaluate 21 representative MLLMs on MVDriveBench, including both open-source and

closed-source models. The open-source group contains 17 models, such as InternVL3-78B (Zhu et al., 2025), Qwen2.5-VL-72B (Bai et al., 2025b), and Cosmos-Reason1-7B (NVIDIA et al., 2025), which are deployed locally using vLLM (Kwon et al., 2023) for efficient inference. The closed-source group contains four advanced proprietary models, including GPT-4o (OpenAI et al., 2024), GPT-5 (Singh et al., 2026), Gemini 2.5 Pro (Comanici et al., 2025), and Gemini 3 Pro (Google DeepMind, 2025), which are evaluated through their official cloud APIs. The complete list of evaluated models and deployment details is provided in Table 6.

D Case Study

To further understand the behavioral differences among evaluated models, we analyze their responses across different question types, with representative examples shown in Figures 18–21. These examples cover *Key Object Identification*, *Critical Traffic Events*, and *Cross-view Spatio-Temporal Consistency*, allowing us to compare how different models interpret visual evidence, associate objects across camera views, and produce final answers.

You are an autonomous driving scene analysis expert. Please carefully analyze this 20-second front-view video (30Hz frame rate) and determine whether the current scenario requires surround-view (multi-view) capability to ensure driving safety.

Analysis Steps

Step 1: Analyze the Ego Vehicle's Current Action
Please observe and determine what the ego vehicle is currently doing:
- Is it driving straight?
- Is it changing lanes or preparing to change lanes?
- Is it turning (left/right)?
- Is it merging onto or exiting a main road?
- Is it passing through an intersection?
- Is it stopping or starting?
- Is it reversing or making a U-turn?
- Other actions?

Step 2: Analyze the Surrounding Environment
Please carefully observe other traffic participants in the video:
- Are there other vehicles around?
- Is there a vehicle following closely behind the ego vehicle (at a very short distance)?
- Is there a vehicle traveling alongside the ego vehicle?
- Is there a vehicle overtaking the ego vehicle or being overtaken by the ego vehicle?
- Are there pedestrians, bicycles, motorcycles, etc.?
- What is the traffic density (sparse / moderate / dense)?

Step 3: Determine Whether Surround-View Capability Is Needed
Based on the above analysis, please determine: **Is the front-view video alone sufficient to ensure driving safety?**

Typical scenarios requiring surround-view capability (answer "Yes"):
1. **Lane changing**: The ego vehicle is changing lanes or preparing to change lanes, and needs to confirm whether there are vehicles in the side-rear blind spots
2. **Close following**: There are vehicles following closely alongside or behind the ego vehicle, requiring continuous monitoring
3. **Overtaking / Being overtaken**: The ego vehicle is overtaking another vehicle or being overtaken by another vehicle
4. **Complex intersections**: When passing through an intersection, it is necessary to observe oncoming cross-traffic
5. **Merging / Exiting**: When merging onto a main road or exiting from a main road, it is necessary to observe side-rear traffic
6. **Narrow roads**: The road is narrow, and it is necessary to confirm whether there are obstacles or oncoming vehicles on both sides
7. **Reversing / U-turn**: It is necessary to observe the surroundings of the vehicle
8. **Other situations requiring blind-spot observation**

Typical scenarios not requiring surround-view capability (answer "No"):
1. **Highway straight driving**: The road ahead is clear, no lane-change is needed, and there are no vehicles in close proximity
2. **Driving on an open road**: There are no other traffic participants around; only the front needs to be monitored
3. **Following a lead vehicle**: Only following the vehicle ahead is needed, with no lane-change or turning requirements, and no vehicles alongside
4. **Other scenarios where monitoring only the front is sufficient to ensure safety**
...

Figure 8: Prompt Template of Clip Selection for Surround-View Necessity.

You are an expert in surround-view spatiotemporal scene understanding and driving reasoning for autonomous driving. Your input is a 5-second, 30 Hz six-view concatenated surround-view video. You are skilled at scene understanding, traffic element understanding, text recognition, cross-view spatial association, target identification and localization, spatiotemporal consistency analysis, target tracking, ego-vehicle behavior analysis, and motion relationship analysis based on multi-view temporal visual information. Please answer strictly based on the video content and subsequent instructions; do not speculate about unobserved information; clearly state uncertainty when confirmation is not possible.

1. [Surround-View Six-View Layout and Spatial Topology Rules]

1. [Input Layout]

The input video is a 2-row × 3-column six-view concatenated frame.
This frame is synchronously composed of 6 real physical cameras; it is not an ordinary 2D planar mosaic.

The layout of the six sub-frames is as follows:

Row 1 - Left: Front Left
Row 1 - Center: Front (wide-angle front view)
Row 1 - Right: Front Right
Row 2 - Left: Left Rear
Row 2 - Center: Rear (telephoto rear view)
Row 2 - Right: Right Rear

2. [Interpretation Principles]

The six-view relationships must be understood according to real physical spatial topology, not merely by 2×3 grid planar adjacency.
Unless otherwise specified, all "left / right / up / down / lower-left / lower-right" in the text refer to the image directions within the corresponding sub-frame, not the grid layout directions.
All "front-left / front-right / left-side / right-side / left-rear / straight-rear / right-rear" in the text refer to physical spatial directions in the ego vehicle's coordinate system.
Rear is the ego vehicle's primary rear-facing viewpoint and is a telephoto rear camera. It is not an ordinary wide-angle rear view; its frame typically has a narrower field of view, stronger long-range observation capability, and a more noticeable scale difference from adjacent rear-facing views.
For the Rear frame, the image left or right side cannot be directly equated with the ego vehicle's left or right side; its physical correspondence must be based on the following topological continuity rules.
When performing cross-view target association, spatiotemporal consistency analysis, target tracking, and motion judgment, the focal-length differences between Rear and Left Rear / Right Rear that may cause apparent scale changes must be considered; one cannot simply judge a target as closer or undergoing a sudden change just because it appears larger or clearer in Rear.
For rearward target analysis, Rear is prioritized for straight-rear and mid-to-long-range rear observation; Left Rear / Right Rear are more suitable for supplementing side-rear and adjacent-area continuous observation.

[Key Topological Continuity Relationships]

The following "↔" indicates that the corresponding regions of two sub-frames are continuous in real physical space:

Rear's image right side ↔ Left Rear's image left side
→ Corresponds to the continuity of the ego vehicle's left-rear space
Rear's image left side ↔ Right Rear's image right side
→ Corresponds to the continuity of the ego vehicle's right-rear space
Front's image left-lower region ↔ Front Left's image right side
→ Corresponds to the continuity of the ego vehicle's left-front space
Front's image right-lower region ↔ Front Right's image left side
→ Corresponds to the continuity of the ego vehicle's right-front space
Left Rear's image right-lower region ↔ Front Left's image left side
→ Corresponds to the continuity of the ego vehicle's left-side space
Right Rear's image left-lower region ↔ Front Right's image right side
→ Corresponds to the continuity of the ego vehicle's right-side space
...

Figure 9: Prompt Template of Scenario Understaning.

Task	Question Format	Example Question
Dynamic Motion Understanding		
Rel. Mot.	Closed-ended short answer	During 0.0–1.5 seconds, what is the relative motion relationship between the ego vehicle and the oncoming silver sedan?
Obj. Mot.	Closed-ended short answer	In the Front Right view, from 2.0 to 4.0 seconds, what is the motion direction of the white FedEx truck relative to the ego vehicle?
Tracking	Closed-ended short answer	A silver sedan appears in the Front Left view at 1.0 second with a bounding box roughly located at [500, 400, 1000, 800]. In which view does this vehicle mainly appear at 3.0 seconds?
Ego Mot.	Closed-ended binary	Does the ego vehicle remain completely stationary throughout the entire 5-second video clip?
Multi-view Scene Understanding		
Intent	Closed-ended multiple choice	Based on the state of the white FedEx truck in the video, what is its most likely intention? A. Overtaking B. Parking or roadside operation C. Reversing D. Preparing to turn left
Cross-view	Closed-ended binary	Is the silver sedan shown in the Front Left view at 1.0 second the same object as the vehicle shown in the Left Rear view at 3.0 seconds?
Key Obj.	Open-ended	Identify the main obstacle that restricts the ego vehicle's movement within the first 2 seconds of the video and describe its location.
Key Event	Open-ended	Describe the main driving maneuver performed by the ego vehicle from 2.0 to 4.0 seconds.
Risk	Closed-ended binary	When the ego vehicle overtakes the FedEx truck, is there a risk of collision with the silver SUV parked on the right side of the road?
Driving Reasoning		
Planning	Closed-ended multiple choice	At 1.0 second, facing a stationary truck ahead and an oncoming vehicle, what is the best decision for the ego vehicle? A. Accelerate and overtake immediately B. Remain stopped and wait for the oncoming vehicle to pass C. Steer right to bypass the truck D. Reverse away from the road segment
Counterfactual	Closed-ended multiple choice	If the ego vehicle changed left to overtake at 1.0 second without waiting for the oncoming silver sedan to pass, what would most likely happen? A. Safely complete the overtake B. Collide head-on with the oncoming sedan C. Rear-end the FedEx truck D. Hit the parked vehicle on the right
Rationality	Open-ended	Explain why the ego vehicle chooses to stop and wait at the beginning of the video instead of overtaking immediately.

Table 5: Task taxonomy and representative question examples in MVDriveBench.

You will receive:
1. A six-view concatenated surround-view video
2. A scene understanding result from the upstream stage (for internal reference only; its existence must not be revealed in the final output)

External Knowledge:
...

Task:
Based on the video, generate 12 high-difficulty VQA question-answer pairs for evaluating the model's multi-step visual reasoning capability.

==========
I. Core Principles
==========

1. First independently observe based solely on the video, then use the upstream results as a hidden reference for internal cross-checking.
2. Upstream results can only be used internally for:
   - Reminding you to re-examine certain targets, events, or time periods
   - Helping you discover visual evidence worth deeper investigation in the video
   - Performing consistency checks
...
==========
II. 12-Question Fixed Coverage and Question Type Ratio
==========

A total of 12 question-answer pairs are output, 1 per category, with question types strictly corresponding as follows:

1. Cross-view spatiotemporal consistency — Closed-ended binary
2. Target motion — Closed-ended short-answer
3. Target tracking — Closed-ended short-answer
4. Ego-vehicle motion — Closed-ended binary
...

==========
III. Internal Workflow (Do Not Output)
==========

1. First extract sufficient spatiotemporal evidence based solely on the video, focusing on:
   - Time-segment changes
   - Cross-view continuity
   - Entering/leaving field of view
   - Occlusion
   - Scale changes
   - Relative position changes
   - Environmental constraints (lane lines, guardrails, curbs, parked vehicles, lighting, weather, etc.)
...

==========
IX. Output Format
==========

...

Figure 10: Prompt Template of QA Construction.

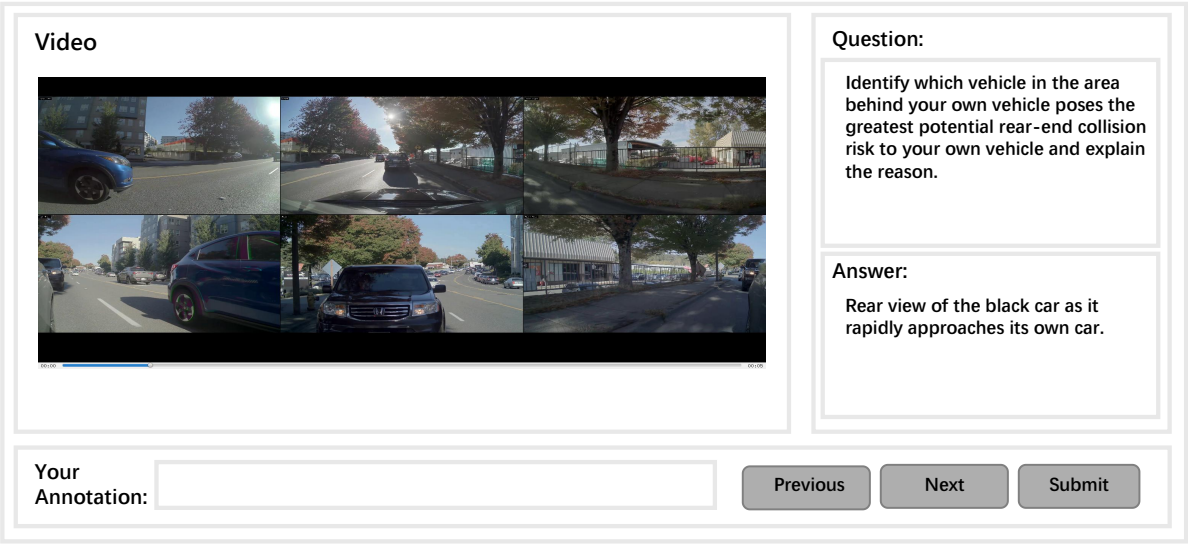


Figure 11: Screenshot of Manual Annotation.

Please answer according to the following question format requirements:

{question_format}

Answer guidelines:

Closed-ended - Binary: The answer field must be "Yes" or "No"
Closed-ended - Short answer: The answer field must be a brief answer (typically 1-5 words)
Closed-ended - Given options: The answer field must output only a single option letter, without the option content (e.g., A / B / C / D)
Open-ended: The answer field must be free text that is concise, direct, and supported by evidence from the video
Please strictly follow the above format requirements. Output only the answer itself, without adding any explanation or extra content.

Question: {question}

Figure 12: Prompt Template of Surround-View Reasoning during Evaluation.

Please answer according to the following question format requirements:

{question_format}

Answer guidelines:

Closed-ended - Binary: The answer field must be "Yes" or "No"
Closed-ended - Short answer: The answer field must be a brief answer (typically 1-5 words)
Closed-ended - Given options: The answer field must output only a single option letter, without the option content (e.g., A / B / C / D)
Open-ended: The answer field must be free text that is concise, direct, and supported by evidence from the video
Note: The current video only contains the front view (Front View) perspective. It is not possible to observe information from other perspectives such as side views or rear views. If the question requires combining other perspectives to answer, you may answer closed-ended - short answer and open-ended questions based on your imagination; for closed-ended - binary and closed-ended - given options questions that require other perspectives, you may guess Yes or No or the option.

Please strictly follow the above format requirements. Output only the answer itself, without adding any explanation or extra content.

Question: {question}

Figure 13: Prompt Template of Front-View Reasoning in Ablation Study.

Please carefully observe the following surround-view video and provide a detailed description of its content.

Description requirements:

Cover key information from all visible perspectives (front view, rear view, front-left view, front-right view, rear-left view, rear-right view)
Describe the ego vehicle's motion state (stationary, moving, turning, etc.)
Describe visible surrounding objects such as vehicles, pedestrians, traffic signs, and their approximate positions
Describe the relative motion relationships between objects
Describe road environment features (number of lanes, intersections, traffic lights, etc.)
Please describe in a single complete paragraph:

Figure 14: Caption Prompt Template of SurVid2txt Reasoning in Ablation Study.

Please answer the question based solely on the following text description of the surround-view video. Do not use any external knowledge or guesswork.

Surround-view video description:
{caption}

Please answer according to the following question format requirements:

{question_format}

Answer guidelines:

Closed-ended - Binary: The answer field must be "Yes" or "No"
Closed-ended - Short answer: The answer field must be a brief answer (typically 1-5 words)
Closed-ended - Given options: The answer field must output only a single option letter, without the option content (e.g., A / B / C / D)
Open-ended: The answer field must be free text that is concise, direct, and supported by evidence from the description
Please strictly follow the above format requirements. Output only the answer itself, without adding any explanation or extra content.

Question: {question}

Figure 15: Inference Prompt Template of SurVid2txt Reasoning in Ablation Study.

```
You are a professional evaluation assistant responsible for assessing the quality of a multimodal model's answers to surround-view scene understanding questions.

## Task Description
Please evaluate the quality of the model's answer based on the following information:
- **Question**: The original question
- **Question Type**: The type of the question (Important Object Recognition / Important Traffic Event / Behavior Rationality Argumentation / Other)
- **Question Format**: The format of the question (Open-ended / Closed-ended - Short Answer)
- **Reference Answer**: The human-annotated correct answer
- **Model Answer**: The answer generated by the multimodal model

## Evaluation Rules

### I. Closed-ended - Short Answer
**Evaluation Method**: Semantic judgment via API call
**Score**: 0 or 1 (no intermediate scores)

**Important: Please evaluate leniently. The following cases should receive a score of 1**:
- Synonym substitution
- Word order variation with identical meaning
- Different phrasing but consistent core information
- Near-synonym substitution

**Scoring Criteria**:
- **1 point**: The model answer is semantically equivalent to the reference answer (as long as the core meaning is consistent, award 1 point; do not be overly strict)
- **0 points**: The model answer contradicts the reference answer, is incorrect, or is completely irrelevant

**Examples**:
- Reference: "NO" → Model: "No" → 1 point (semantically equivalent)
- Reference: "Driving in the same direction" → Model: "Driving in the same lane" → 1 point (semantically equivalent)
- Reference: "Rear" → Model: "Rear-View" → 1 point (semantically equivalent)
- Reference: "more quickly" → Model: "speed is bigger" → 1 point (semantically equivalent)
- Reference: "Yes" → Model: "No" → 0 points (opposite meaning)
- Reference: "forward" → Model: "move backward" → 0 points (opposite meaning)

---

### II. Open-ended Questions (0-1 scoring, in increments of 0.1)

#### 1. Important Object Recognition — Open-ended
Evaluate the model's ability to identify and describe key objects in the scene.

**Scoring Criteria:**
| Score | Criterion |
|-------|-----------|
| 1.0 | Perfect recognition — accurately identifies object type, color, and salient features, fully consistent with the reference answer |
| 0.9 | Excellent — accurately identifies object type and main features, with only minor differences |
| 0.7-0.8 | Good — correctly identifies object type and captures some key features, but omits important details |
| 0.5-0.6 | Fair — identifies object type but feature descriptions are vague or incomplete, or contain minor errors |
| 0.3-0.4 | Poor — object type is correct but most feature descriptions are wrong, or key identifying information is missing |
| 0.1-0.2 | Very Poor — object type is incorrectly identified or the answer is irrelevant to the question |
| 0.0 | Completely incorrect or no answer |

**Examples**:
- Reference: A large white van with pink graffiti on the side
- Model A: A large white truck with pink patterns on the side → 0.9 points (correct identification, "graffiti" described as "patterns")
- Model B: A white truck → 0.5 points (type correctly identified, but missing key features)
- Model C: A black sedan → 0.0 points (completely incorrect)

---
```

Figure 16: Prompt Template of LLM-as-Judge Evalutation Protocol.

2. Important Traffic Event — Open-ended
Evaluate the model's ability to understand and describe the main events in a traffic scene.

Scoring Criteria:
Score	Criterion
1.0	Perfect description — accurately describes the event subject, action, and direction/state, semantically fully consistent with the reference answer
0.9	Excellent — accurately describes the core event; different phrasing but semantically equivalent
0.7–0.8	Good — describes the main event but details are incomplete (e.g., missing direction, state, etc.)
0.5–0.6	Fair — captures partial information about the event, but the description is vague or contains minor errors
0.3–0.4	Poor — most of the event description is inaccurate, or key information is missing
0.1–0.2	Very Poor — event understanding is incorrect or the answer is irrelevant to the question
0.0	Completely incorrect or no answer

Examples:
- Reference: Congested slow-moving traffic with small gaps between vehicles moving slowly
- Model A: Traffic congestion, vehicles moving slowly → 0.9 points (core semantics consistent)
- Model B: Many vehicles → 0.5 points (captures partial information, but description is overly brief)
- Model C: Road is clear → 0.0 points (completely incorrect)

---

3. Behavior Rationality Argumentation — Open-ended
Evaluate the model's ability to reason about and argue for the rationality of driving behaviors.

Scoring Criteria:
Score	Criterion
1.0	Perfect argumentation — accurately identifies key factors, provides a complete and reasonable causal explanation consistent with the core logic of the reference answer
0.9	Excellent — argumentation logic is correct, key factors are accurately identified; different phrasing but equivalent reasoning
0.7–0.8	Good — argumentation direction is correct, identifies some key factors, but reasoning is incomplete
0.5–0.6	Fair — provides some reasonable explanation, but key factors are inaccurately identified or reasoning has gaps
0.3–0.4	Poor — argumentation logic is weak, or only provides a conclusion without reasonable explanation
0.1–0.2	Very Poor — argumentation is incorrect or the answer is irrelevant to the question
0.0	Completely incorrect or no answer

Examples:
- Reference: Because there are large vehicles on both sides (a red truck on the left, a graffiti van on the right) in close proximity, leaving insufficient safe lane-change space
- Model A: Surrounding vehicles are dense, insufficient space for lane change → 0.8 points (correct argumentation, but does not specifically identify the key vehicles)
- Model B: Because there is a car ahead → 0.3 points (argumentation is overly brief, fails to capture key factors)
- Model C: Because the weather is nice → 0.0 points (completely irrelevant)

Output Format

Please output the evaluation result strictly in the following JSON format:
```json
{{
    "score": a decimal number between 0.0 and 1.0 (one decimal place),
    "reason": "brief explanation of the scoring rationale, noting the strengths and weaknesses of the answer"
}}
```

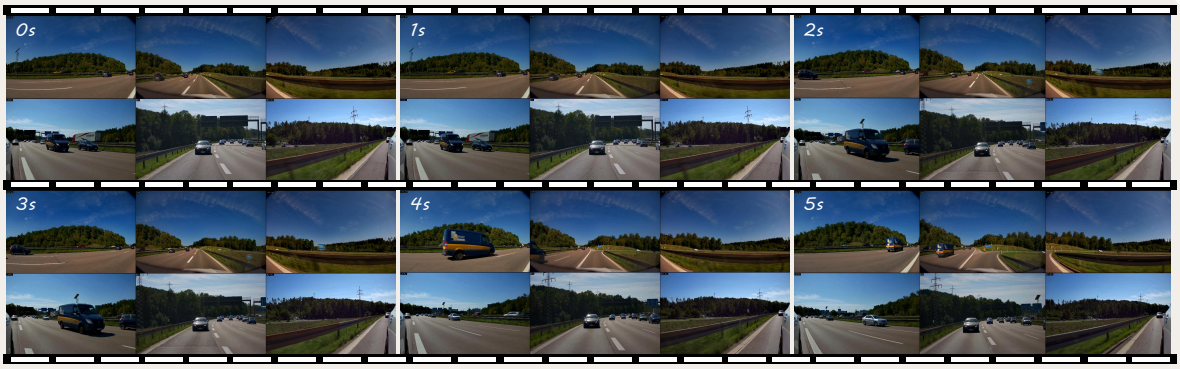
Notes:
- For Closed-ended - Short Answer, the score must be either 0 or 1 (evaluate leniently; award 1 point for semantic equivalence)
- For Open-ended questions, the score can be any value from 0.0, 0.1, 0.2, ..., 1.0

Input Data

Question Type: {question_type}
Question Format: {question_format}
Question: {question}
Reference Answer: {original_answer}
Model Answer: {model_answer}

Please evaluate and output the JSON result:

Figure 17: Prompt Template of LLM-as-Judge Evalutation Protocol.



Q: Please describe the characteristics and status of the vehicle directly behind your own vehicle in the Rear view.
A: A black car follows the car at a safe distance.
R: A black car is always shown in the center of the Rear view in the video. The front of the vehicle is facing towards its own vehicle, indicating that it is traveling in the same direction. During the entire 5-second video, its size in the picture remains basically unchanged, indicating that its relative distance to its own vehicle is stable and it is in a following state.
GPT-5: A black car (medium-sized/sedan) is following in the center directly behind the same lane, with daytime running lights/headlights on; there is no turn signal or sign of lane change, the following distance is medium, the speed is similar to the own car, and the state is stable.
ReCogDrive-VLM-8B: The vehicle directly behind your own car was a silver car. It was at a moderate distance and maintained a safe following distance. It was traveling at a stable speed and showed no signs of accelerating or decelerating.
Cosmos-Reason1-7B: Rear view of the vehicle directly behind the ego vehicle
Cosmos-Reason2-8B: A silver Maserati car was driving directly southwest of the own car, and it was switching to the traction line where the own car was.
MiMo-Embodied-7B: <think>\nThe user now needs to describe the characteristics and status of the vehicle directly behind the own vehicle in the Rear view. Looking at the Rear screen, the vehicle directly behind should be in a driving state. What are its characteristics such as color, type, etc.? In the rear view of the picture, the vehicle directly behind is driving normally, so it should be described briefly. Need to confirm: In the Rear view, the vehicle directly behind your own vehicle is in a driving state. For example, may it be a certain type of vehicle? Looking at Rear's perspective in the screen, the vehicle directly behind him should be following the car normally. Therefore, it is described as "the vehicle directly behind is driving normally"
MiMo-VL-7B-SFT: <think>\nThe user now needs to describe the characteristics and status of the vehicle directly behind the own vehicle in the Rear view. Looking at the Rear screen, the vehicle directly behind should be driving at a safe distance. What are the characteristics such as car model, color, etc.? But first look at the Rear perspective in the video. The vehicle directly behind your own car should be following the car normally. So describe it concisely. \n\nFirst look at the Rear view. The vehicle directly behind your own vehicle is driving at a safe distance from the screen. The vehicle characteristics may be a certain type of vehicle, but the more accurate status is: keep the distance between vehicles and drive normally. So the answer should be concise. \n</think>The vehicle directly behind your vehicle keeps a safe distance and drives normally.

Figure 18: Responses of different models to Key Object Identification in Clip 3d820f09-db1a-4dbc-bc98-10da9bc7827b.

Q: Please describe the type and color of the vehicle parked at the market entrance to the right of your vehicle.
A: A blue retro truck
R: The Front Right view clearly shows the scene in front of the right side of your vehicle. In the sidewalk area, parked is a truck with an old-fashioned design. The vehicle's body color is distinctly light blue or cyan. Its exterior features such as round headlights and grille indicate that it is a retro model.
GPT-5: **White SUV**
ReCogDrive-VLM-8B: The vehicle parked at the entrance of the market on the right front of the vehicle is white.
Cosmos-Reason1-7B: **SUV, blue**
Cosmos-Reason2-8B: Parked at the entrance of the market to the right of your car is a light blue flatbed truck.
MiMo-Embodied-7B: <think>\nThe user now needs to describe the vehicle at the market entrance on the right front. Look at the car at the front right of the screen at the entrance to the market. It's a light blue pickup truck. To confirm location and vehicle type color. So the answer should be concise. \n</think>\n\nLight blue pickup truck
MiMo-VL-7B-SFT: <think>\nThe user now needs to look at the vehicle at the market entrance on the right front. Looking at the Front Right screen, the car at the entrance of the market is light blue (turquoise) in color and is a pickup truck.</think>a blue pickup truck.

Figure 19: Responses of different models to Key Object Identification in Clip 3c2bf689-da99-44a5-95ad-da0b1e306ba8.

Q: What is the main driving behavior being performed by the car at the end of the video?
A: Stop at intersection
R: In the second half of the video, the speed of the vehicle dropped significantly, and it finally stopped in front of the road stop line. There are clear parking control facilities at the intersection ahead, such as pavement text and sign posts. This series of actions constitutes a complete intersection parking process. Therefore the main driving behavior is intersection parking.
GPT-5: Stop and wait at the intersection stop line (do not proceed before the Stop sign).
ReCogDrive-VLM-8B: The vehicle is coming to a stop at an intersection.
Cosmos-Reason1-7B: parking
Cosmos-Reason2-8B: Wait at the intersection
MiMo-Embodied-7B: <think>\nUsers now need to answer open-ended questions and watch their car's behavior at the end of the video. Looking at the picture, your car should be slowing down and stopping because there is a stop sign, so you need to confirm. Then organize the language. \n</think>\n\nThe vehicle is slowing down and stopping.
MiMo-VL-7B-SFT: <think>\n</think>\n\nParking.

Figure 20: Responses of different models to Critical Traffic Events in Clip 4dfd1641-6beb-49d9-a4a7-764a84720f65.

Q: Is the approaching vehicle shown in the Left Rear view at 1s in the video the same target as the overtaking vehicle on the left shown in the Front Left view at 2.0s?
A: Yes
R: The Left Rear view in the video shows a dark vehicle approaching quickly from the left rear at 1 seconds. The Front Left view then shows a dark SUV overtaking the self-car from the left at 2.0s. According to the spatial topological continuity, the vehicle on the left rear moving forward will naturally enter the left front field of view, and the time sequence is consistent with the vehicle characteristics. Therefore, the target in both views is the same object and the answer is Yes.
GPT-5: Yes.
ReCogDrive-VLM-8B: No
Cosmos-Reason1-7B: yes
Cosmos-Reason2-8B: No
MiMo-Embodied-7B: <think>\nThe user now needs to determine whether the vehicles in the two views are the same target. Check whether the vehicle that Left Rear approaches at 1s and the vehicle that Front Left overtakes on the left at 2.0s are the same vehicle. From the video, it should be the same target, so answer Yes. \n</think>\n\nYes.
MiMo-VL-7B-SFT: <think>\n</think>\n\nYes.

Figure 21: Responses of different models to Cross-view Spatio-Temporal Consistency in Clip 701526ce-7268-48b0-8e0b-5a63f718c92f.

Model	#Frames	Imple.	max_tokens	temperature	top_p	top_k
Closed Source Models						
GPT-5	10	API	4096	0.1	–	–
GPT-4o	10	API	4096	0.1	–	–
Gemini-2.5-Pro	10	API	4096	0.1	–	–
Gemini-3-Pro	10	API	4096	0.1	–	–
Open Source MoE Models						
Qwen3.5-397B-A17B	10	vLLM	4096	0.1	0.8	40
ERNIE-4.5-VL-28B-A3B-Thinking	10	vLLM	4096	0.1	0.8	40
Qwen3-VL-30B-A3B-Thinking	10	vLLM	4096	0.1	0.8	40
Open Source Models $>$ 7B						
InternVL3-78B	10	vLLM	4096	0.1	0.8	40
Qwen2.5-VL-72B	10	vLLM	4096	0.1	0.8	40
InternVL3-38B	10	vLLM	4096	0.1	0.8	40
Qwen2.5-VL-32B	10	vLLM	4096	0.1	0.8	40
Qwen3.5-27B	10	vLLM	4096	0.1	0.8	40
GLM-4.1V-9B-Thinking	10	vLLM	4096	0.1	0.8	40
Video-R1-7B	10	vLLM	4096	0.1	0.8	40
LLaVA-OneVision-Qwen2-7B-OV	10	vLLM	4096	0.1	0.8	40
VGLM-QA-VGGT-8B	10	HF	4096	0	–	–
Specialist Models						
ReCogDrive-VLM-8B	10	vLLM	4096	0.1	0.8	40
Cosmos-Reason1-7B	10	vLLM	4096	0.1	0.8	40
Cosmos-Reason2-8B	10	vLLM	4096	0.1	0.8	40
MiMoEmbodied-7B	10	vLLM	4096	0.1	0.8	40
MiMo-VL-7B-SFT	10	vLLM	4096	0.1	0.8	40

Table 6: Inference configurations of evaluated MLLMs.