WaymoQA: A Multi-View Visual Question Answering Dataset for Safety-Critical Reasoning in Autonomous Driving

Seungjun Yu[1], Seonho Lee[2], Namho Kim[3], Jaeyo Shin[1],
Junsung Park[1], Wonjeong Ryu[1], Raehyuk Jung[1], Hyunjung Shim[1]*

[1]KAIST, [2]KRAFTON, [3]Hanyang University, Seoul, Republic of Korea

seungjunyu@kaist.ac.kr, glanceyes@krafton.com, knhnh@hanyang.ac.kr, jaeyo_shin@kaist.ac.kr
jshackist@kaist.ac.kr, petac@kaist.ac.kr, mulnyangi@naver.com, kateshim@kaist.ac.kr

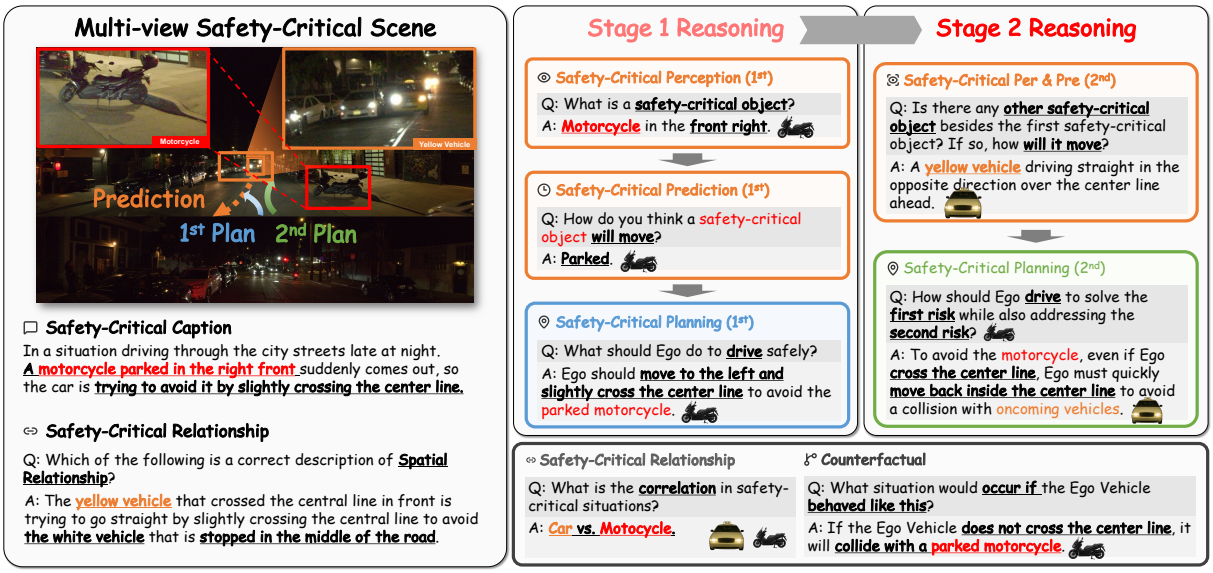


Figure 1: WaymoQA overview of action-induced two-stage safety-critical reasoning. A Stage 1 maneuver resolves the primary hazard but induces a secondary risk, requiring Stage 2 planning under the changed risk landscape.

Abstract

In real-world driving, resolving a primary hazard often precipitates a secondary one: swerving to avoid a stalled vehicle may force the ego car into oncoming traffic. Existing driving QA datasets fail to capture this cascading causality, evaluating perception, prediction, and planning in silos without anticipating the risks action induces. We introduce WaymoQA, a multi-view dataset that frames safety-critical reasoning as a two-stage task: resolving the immediate threat, then mitigating the action-induced secondary risk. WaymoQA provides 35,000 human-annotated QA pairs across image and video modalities, with 28K training and 6.4K test examples. Experiments show that SOTA models struggle on safety-critical scenes, while fine-tuning on WaymoQA improves both reasoning accuracy and consistency. WaymoQA thus serves as a rigorous benchmark and resource for safety-aware autonomous driving.

Our code and data are provided in https://github.com/sjyu001/WaymoQA.

1 Introduction

Multimodal Large Language Models (MLLMs) have rapidly advanced from general vision-language tasks to autonomous driving, they now support end-to-end planning, scenario generation, and vision-language action modeling (Feng et al., 2025; Fu et al., 2024; Guo et al., 2025; Huang et al., 2024; Ma et al., 2024; Shao et al., 2024a). Within this landscape, Driving Question Answering (Driving QA) has emerged as a promising framework for evaluating high-level scene understanding and decision-making (Marcu et al., 2024; Li et al., 2025b), since QA format directly exposes what a model perceives, predicts, and decides.

Driving QA datasets have evolved from single-view setups such as DRAMA (Malla et al., 2023) and DriveLM (Sima et al., 2024) to multi-camera

*Corresponding Author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 21721–21744
October 24-29, 2026 ©2026 Association for Computational Linguistics

frameworks like NuScenes-QA (Qian et al., 2024) and NuPlan-QA (Park et al., 2025), progressively improving spatial coverage and inter-agent modeling. However, both generations share a common blind spot: they are overwhelmingly built from normal driving scenarios (Fu et al., 2025; Li et al., 2024b; Liu et al., 2024c; Zeng et al., 2025), leaving safety-critical events, the rare but consequential moments that most challenge real-world deployment, largely untested.

Recent benchmarks have begun addressing this gap. NAVSAFE (Sima et al., 2025), DVBench (Zeng et al., 2025), and VRU-Accident (Kim et al., 2025) introduce safety-critical evaluation for driving VQA, yet three key limitations persist. (**L1**) They are mainly evaluation-only and do not provide training data for learning safety-critical reasoning. (**L2**) They often rely on limited or single-view inputs, which can miss blind spots and occluded agents; and (**L3**) They treat perception, prediction, and planning as independent questions, without evaluating action-induced downstream risks where resolving one hazard can create another. As shown in Fig. 2, the same scene can lead to opposite decisions depending on view coverage: a lane change appears feasible from the front view, but multi-view input reveals a rear-side vehicle that would turn that maneuver into a collision. This illustrates that safety-critical reasoning requires both multi-view evidence and explicit reasoning over how a chosen action reshapes the surrounding risk landscape.

To address these limitations, we introduce WaymoQA, the first training-enabled, safety-critical multi-view driving QA dataset. WaymoQA is built from rare pre-crash scenarios in the Waymo End-to-End (Waymo E2E) dataset (**?**) and filtered using the 47 U.S. NHTSA pre-crash scenario types (**?**). Here, *long-tail* specifically refers to rare pre-crash scenario types defined by the NHTSA taxonomy, rather than rare environmental conditions such as adverse weather, nighttime driving, or heavy traffic. WaymoQA contains 35,000 human-written question-answer pairs across safety-critical videos and manually selected key frames. It provides 28,585 training examples and 6,415 test examples, supporting both open-ended QA and multiple-choice formats. Each scene is observed through eight synchronized camera views covering front, side, and rear directions, supplying the surrounding context that single-view setups miss.

Central to WaymoQA is its formulation of *Safety-Critical Reasoning* as a compositional two-stage task. Stage 1 resolves the immediate risk by identifying the safety-critical object, predicting its behavior, and selecting a feasible safe action. Stage 2 then addresses the secondary risks induced by the Stage 1 action, reflecting real-world driving where mitigating one hazard can introduce another. For example, as shown in Fig. 1, the ego vehicle first avoids a parked motorcycle by briefly crossing the center line, but must then return to its lane before colliding with an oncoming vehicle. WaymoQA further combines two complementary QA modalities: Video QA captures realized temporal events, while Image QA supports decision-oriented and counterfactual reasoning from key frames, broadening the space of possible actions and outcomes beyond the single realized trajectory.

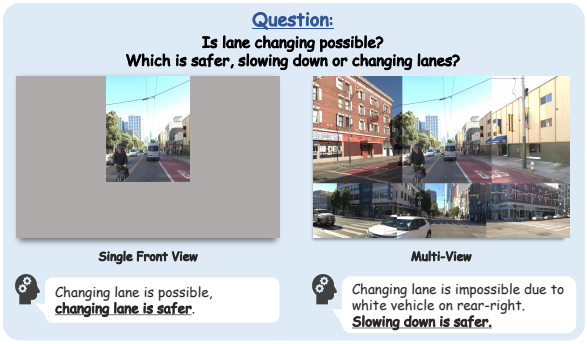


Figure 2: **Same-Scene View Comparison.**

Comprehensive experiments show that pre-trained MLLMs consistently struggle with safety-critical reasoning, underperforming markedly on safety-critical questions compared to normal scenes. Fine-tuning on WaymoQA improves ImageQA, VideoQA, open-ended QA, and two-stage consistency, indicating gains beyond answer-option selection toward free-form and decision-conditioned reasoning. Transfer, ablation, and robustness analyses further show that WaymoQA supervision improves external driving QA benchmarks and planning metrics, benefits from multi-view inputs, and remains stable under answer-order permutation and environmental variations.

Our contributions are threefold: (1) we define multi-view safety-critical reasoning for autonomous-driving VQA; (2) we introduce WaymoQA, the first training-enabled safety-critical multi-view driving QA dataset with image/video and open-ended/MCQ formats; and (3) we present MCQ, consistency, transfer, ablation, and robustness evaluations showing that WaymoQA supervision improves safety-critical reasoning.

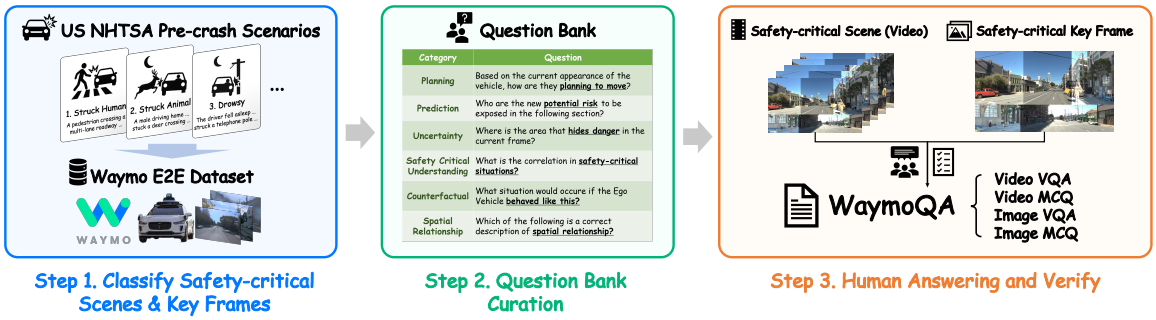


Figure 3: **WaymoQA construction pipeline.** We filter safety-critical Waymo E2E scenes, curate a structured QA bank, and perform human answering and verification to produce Video/Image VQA and MCQ splits.

2 Related Works

MLLM in Autonomous Driving. MLLMs have shown strong high-level reasoning across diverse tasks (Liu et al., 2024d; Stone et al., 2023; Chen et al., 2024a; Hong et al., 2024), motivating their use in autonomous driving (Chen et al., 2024c; Cui et al., 2024; Zhou et al., 2025c). Recent studies extend MLLMs from perception and captioning to driving-oriented reasoning, planning, and trajectory generation, including spatial/action reasoning (Tian et al., 2024; Xu et al., 2024), high-resolution or 3D-enhanced perception (Ding et al., 2025; Han et al., 2025), generative planning (Hu et al., 2023; Mao et al., 2023), agentic driving (Jin et al., 2024), VLA models (Zhou et al., 2025b), and text-to-trajectory conversion (Hwang et al., 2024). Overall, the field is shifting from scene narration toward decision-making coupled with perception and control.

Safety-Critical Autonomous Driving. Safety-critical driving is difficult due to rarity, diversity, and causal dependencies. Prior work has studied planner-driven stress testing (Rempe et al., 2022; Zhong et al., 2022, 2023), adversarial, controllable, and RL-based scenario generation (Xie et al.; Lin et al., 2025; Liu et al., 2024a), real-video anomaly analysis (Yao et al., 2022), safety-focused datasets and benchmarks (Luo and Wang, 2023; Wang et al., 2024), and simulator-scale synthesis (Zhang et al., 2024; Dosovitskiy et al., 2017). Recent efforts further improve realism and multi-view alignment through diffusion, reinforcement learning, and causal rationales (Li et al., 2025a; Zheng et al., 2024; Zhou et al., 2025a; Shao et al., 2024b; Fang et al., 2024).

3 WaymoQA Dataset

We describe WaymoQA, a human-annotated multi-view VQA dataset for safety-critical reasoning. Figure 3 summarizes the construction pipeline.

3.1 Scene Selection and Multi-View Inputs

We build WaymoQA from the Waymo End-to-End (Waymo E2E) dataset (Xu et al., 2025), which contains long-tail real-world driving scenarios. To reduce subjectivity, we select candidate safety-critical scenes using the 47 pre-crash scenario types defined by the U.S. National Highway Traffic Safety Administration (NHTSA) (Administration et al., 2007), providing reproducible criteria for collision-relevant risks, abrupt agent interactions, and hazardous traffic configurations. Final safety-critical/normal classification and key-frame selection are performed by human annotators, without the use of an automatic model.

Each scene has synchronized multi-view observations: Image QA uses eight same-timestamp views, and Video QA uses multi-view clips from the same scenario. This coverage is essential because action feasibility may depend on agents beyond the front view; for example, a lane change may appear safe from the front camera but become unsafe once a rear-side vehicle is visible.

Safety-critical events are brief, often lasting about 2–4 seconds in a 22-second Waymo E2E scenario. Sampling critical moments would overrepresent hazardous states and weaken normal-to-critical evaluation. We therefore curate key frames covering both contexts while avoiding near-duplicates, using a 2:3 normal-to-critical ratio, one-second minimum interval, and at least five key frames per video. This preserves temporal context for Video QA while yielding balanced image-level questions.

To prevent data leakage, we split the dataset at the Waymo scenario-token level, such that all clips, key frames, and QA pairs derived from the same scenario are assigned to a single split. The train, validation, and test splits contain 127, 148, and 54 unique scenarios, respectively, with zero scenario-token overlap across the three splits.

Dataset	Normal	Safety-Critical	Multi View	Train Data	Video	Image	Per.	Nor. Rea.	SC. Rea.	2SS. Rea.	Uncer-tanty	Counter-factual	Amount
NuScenes-QA (Qian et al., 2024)	✓	✗	✓	✓	✗	✓	✓	✗	✗	✗	✗	✗	460K
NuplanQA (Park et al., 2025)	✓	✗	✓	✓	✗	✓	✓	✗	✗	✗	✗	✗	1M
DriveLM (Sima et al., 2024)	✓	✗	✓	✗	✗	✓	✓	✓	✗	✗	✗	✗	443K
DriveMLLM (Guo et al., 2024)	✓	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗	✗	4K
BDD-X (Kim et al., 2018)	✓	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗	✗	26K
DRAMA (Malla et al., 2023)	✓	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗	✗	14K
DriveBench (Xie et al., 2025)	✓	✗	✓	✗	✗	✓	✓	✓	✗	✗	✗	✗	20K
VRU-Accident (Kim et al., 2025)	✓	✓	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗	6K
DVBench (Zeng et al., 2025)	✓	✓	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗	10K
STRIDE-QA (Ishihara et al., 2026)	✓	✗	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	16M
WaymoQA	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**35K**

Table 1: **Comparison of driving QA datasets and reasoning coverage.** Abbreviations: Perception(Per.), Normal Reasoning(Nor.Rea.), Safety-Critical Reasoning(SC.Rea.), Two-Stage Safety-Critical Reasoning(2SS.Rea.).

3.2 Question Bank Curation

Before authoring QA pairs, we construct a **62-item question bank** to standardize coverage across input formats, scene types, and safety-critical reasoning categories. The bank follows a three-level hierarchy of input format, scene type, and reasoning category, covering Caption, Scene Understanding, Perception, Prediction, Planning, Uncertainty, Counterfactual, Special Event, Spatial Relationship, Temporal Grounding, and Traffic Signal. Each template is labeled with structured metadata, with details provided in Appendix B.

Using this curated question bank, WaymoQA contains 35K human-authored QA pairs, split into 28,585 training and 6,415 test examples. The training split contains open-ended QA pairs that provide rich supervision for model fine-tuning. The test split contains human-written four-option multiple-choice questions (MCQs), enabling objective evaluation with a fixed answer key. Table 1 compares WaymoQA with prior driving QA datasets. Unlike previous datasets, WaymoQA jointly provides safety-critical scenarios, multi-view inputs, trainable annotations, image/video QA, counterfactual questions, and two-stage safety-critical reasoning.

3.3 Two-Stage Safety-Critical Reasoning

WaymoQA formulates safety-critical driving QA as a two-stage action-induced reasoning task. Stage 1 asks the model to identify the primary hazard, predict its behavior, and select a feasible action to resolve the immediate risk. Stage 2 is induced by this Stage 1 action, as the initial maneuver may change the ego vehicle's position and interaction with surrounding agents, exposing secondary risks that were not the primary threat before. The model must then choose an action that mitigates the downstream risk while preserving the Stage 1 objective.

Unlike ordinary single-step scene understanding, the Stage 2 answer cannot be inferred from the current scene alone; it depends causally on the Stage 1 maneuver. For example, crossing the center line may be necessary to avoid a parked motorcycle, but this maneuver can induce a new conflict with an oncoming vehicle. Stage 2 evaluates whether the model can anticipate how its action changes the future risk landscape, rather than merely recognizing visible objects or predicting the next frame.

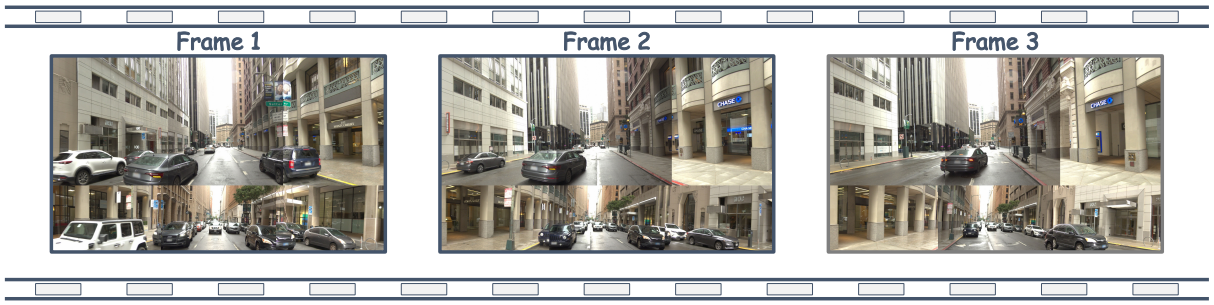
This design separates immediate-risk errors from downstream-foresight errors and supports counterfactual branching, where different Stage 1 actions induce different Stage 2 risks. WaymoQA evaluates not only hazard identification, but also causal reasoning over action-induced safety consequences.

3.4 Video vs. Image QA

Safety-critical scenarios are crucial but rare in practice, making it difficult to achieve broad coverage when constructing reasoning data from video alone. We therefore leverage two complementary perspectives. Video QA enables temporal reasoning about what actually occurred, while Image QA extends safety-critical reasoning by exploring alternative decisions from a single key frame.

In the example of Fig. 4, a vehicle cuts in from the left as the ego changes lanes, and the ego moves to the right lane to avoid a collision. From a single key frame of the same situation, Image QA can ask which actions are feasible at this time, which option is safer given visible risks, and whether a lane change is possible or slowdown is preferable. Since these questions move beyond the single realized path, they expand the set of alternative actions and near term outcomes for the same scene.

Combining the two perspectives increases the diversity of reasoning signals for rare long-tail safety-critical events. This enables more robust learning and evaluation even when scenario data are limited.



Scene Understanding

Video Summary :
Ego vehicle proceeded straight while avoiding a parked vehicle on the left side of a four-lane road in the city.
There was **a vehicle changing to the right lane from the left front road into the road Ego vehicle was driving on**, and since the distance between that vehicle and mine was close, Ego vehicle changed lanes from left to right.

Image Caption :
A situation where Ego vehicle is changing from the second right turn-only lane to the first right turn lane on a four-lane road in the city.
A **gray vehicle in the left front is changing lanes to the right to move in front of the vehicle**.

Video QA	Image QA		
Question : How did the safety-critical object move and how did the ego vehicle behave? **Answer :** The safety-critical object moved in front of the vehicle by changing lanes from left to right, the Ego-vehicle changed lanes to the right twice.	**Question :** What actions are currently possible? **Answer :** Go straight, stop, change to the right lane	**Question :** What should Ego do to drive safely? **Answer :** The ego vehicle should slow down to avoid colliding with the vehicle merging from the left.	**Question :** Is lane changing possible? Which is safer, slowing down or changing lanes? **Answer :** Lane changing is possible, slowing down is safer.
Question : Who is responsible in a safety-critical situation? What is the cause? **Answer :** A gray vehicle changing lanes in the left front did so abruptly without maintaining a safe distance and changed lanes consecutively.	**Question :** If the Ego vehicle behaved like this, what situation might occur? **Answer :** If the Ego vehicle continues straight without slowing down, a collision with a gray vehicle changing lanes ahead will occur.	**Question :** What situation might have occurred if the Ego vehicle had behaved this way? **Answer :** If the Ego vehicle had not slowed down, a collision would have occurred with the vehicle changing lanes.	

Figure 4: Examples of Video/Image QA in WaymoQA.

3.5 Formats and Evaluation

WaymoQA provides two complementary formats. Open-ended QA pairs are used for training, whereas four-option multiple-choice questions (MCQs) are used for validation and testing. Importantly, models are fine-tuned exclusively on open-ended QA pairs and are never exposed to MCQ answer options or distractors during training. Therefore, improvements in MCQ evaluation cannot result from direct memorization of WaymoQA-specific distractor patterns during fine-tuning. We adopt a four-option MCQ design and write all options by humans. The position of the correct answer is uniformly distributed across the four slots to avoid position bias and reduce random guessing. This choice avoids LLM-as-a-judge scoring and enables objective accuracy evaluation with a fixed answer key. Prior work (Chen et al., 2024b; Chen et al.; Wataoka et al., 2024) has shown that LLM-based grading can exhibit biases, such as a preference for longer answers regardless of correctness and a tendency to favor responses written in a style similar to the evaluator. Our MCQ setup is therefore intended to provide a consistent and fair evaluation without relying on an LLM grader.

Statistic	Value
Annotators	7
Independent annotations per scene	≥ 3
Annotation manager flagging rate	26%
Re-annotation threshold	Fleiss' $\kappa < 0.7$
ImageQA κ Normal / Others / Safety-Critical	0.95 / 0.90 / 0.97
VideoQA κ Normal / Others / Safety-Critical	0.90 / 0.91 / 0.98
Independent verifiers	3
Verification flag rate	34% $\rightarrow$ 3% $\rightarrow$ 0%

Table 2: Annotation reliability and verification statistics.

3.6 Human Answering and Verification

Recent studies have begun to automatically generate driving QA data with MLLMs (Khalili and Smyth, 2025; Parikh et al., 2025). However, such data may reinforce pretrained model biases, reduce diversity, and contribute to model collapse (Shumailov et al., 2024). We therefore employ seven Ph.D. annotators with prior knowledge of driving and autonomous-driving research for internal human annotation and verification. Annotation was conducted internally as part of the research project,

Model	ImageQA										VideoQA								
	Normal	Others	Safety-Critical							Overall	Normal	Others	Safety-Critical						Overall
	Overall	Overall	SU	PE	PR	PL	SC	UN	Overall		Overall	Overall	SC	SU	PE	TG	UN	Overall	
Human Experts	96.1	96.3	96.8	98.0	96.4	94.3	98.7	94.9	96.0	96.1	96.9	96.6	97.2	97.2	96.0	94.0	92.5	95.9	96.5
GPT 5.2	75.7	57.0	68.4	65.6	73.8	62.3	57.6	53.0	64.5	67.4	79.1	78.8	54.1	65.5	50.0	49.0	81.3	61.6	72.6
LLaVA1.5(7B)	35.9	37.5	24.5	21.3	24.7	25.5	27.2	17.0	23.2	30.7	42.4	34.7	13.5	27.5	32.6	27.4	33.3	27.5	35.9
LLaVA-OneVision(7B)	66.7	51.7	50.8	42.2	56.9	53.2	45.5	**64.9**	54.4	58.6	65.8	82.1	32.4	43.4	40.3	29.4	62.5	42.3	59.2
InternVL3.5(8B)	40.1	46.6	26.6	30.4	23.7	41.1	25.3	22.2	28.9	37.0	45.1	50.8	29.7	28.2	51.9	19.6	43.7	32.9	41.4
InternVL3.5(4B)	38.0	42.3	21.7	24.1	23.6	39.2	41.1	20.2	27.8	34.4	36.3	58.4	37.8	23.4	42.3	17.6	39.5	29.3	36.7
InternVL3(8B)	33.6	47.0	28.8	24.9	27.1	46.2	41.7	27.8	32.9	35.9	36.3	39.7	40.5	27.5	50.0	25.4	56.2	36.2	36.8
InternVL3(14B)	40.2	55.6	23.4	32.8	52.1	45.3	45.8	39.0	42.1	44.0	46.6	60.4	59.4	27.3	51.9	25.0	57.4	38.6	46.3
Qwen2.5-VL(32B)	71.2	60.3	52.6	44.6	74.3	64.6	56.4	61.4	62.6	65.4	72.3	79.7	43.2	46.2	51.9	31.3	67.8	47.6	64.1
Qwen2.5-VL(7B)	68.1	55.6	54.4	55.3	71.0	64.0	43.0	54.6	61.1	62.4	72.7	64.4	32.4	53.1	55.7	31.3	64.5	49.5	63.8
Phi3.5-vision	48.9	44.5	32.0	25.6	41.8	45.6	44.3	46.5	40.9	44.7	50.9	52.5	37.8	31.7	32.6	24.5	54.1	32.6	45.1
Gemma3(4B)	59.7	48.7	34.2	37.1	53.8	44.3	51.3	33.8	43.5	50.6	60.0	66.1	54.0	29.6	38.5	54.9	54.2	41.1	53.8
Gemma3(27B)	64.3	62.6	67.2	61.6	67.1	59.4	54.4	49.4	60.4	62.3	65.1	69.4	43.2	53.7	55.7	39.2	62.5	51.8	60.7
InternVL2-DA-DriveLM(4B)	54.2	37.7	44.4	24.1	41.2	26.8	38.6	46.1	36.8	43.6	57.0	52.5	37.8	43.4	30.7	27.4	58.3	40.5	50.3
OpenDriveVLA	54.9	43.9	57.1	50.7	68.3	53.1	41.1	49.6	56.0	53.4	67.9	72.0	43.2	49.7	50.0	35.3	54.1	47.5	60.9
LLaVA-OneVision(7B) (FT, Ours)	74.3	62.8	64.7	61.9	67.4	61.9	59.3	59.3	65.5	68.3	71.1	**84.5**	54.2	60.7	59.2	44.1	68.1	58.3	68.1
InternVL3.5(8B) (FT, Ours)	61.6	65.6	57.1	54.1	53.2	50.3	51.3	54.2	53.1	58.7	62.1	76.1	51.3	51.3	60.1	37.2	63.7	52.4	60.4
Qwen2.5-VL(7B) (FT, Ours)	**74.9**	**72.7**	**76.1**	**76.6**	**77.8**	**65.9**	**67.8**	63.8	**71.2**	**72.9**	**76.0**	83.1	**63.5**	**66.2**	**63.5**	**58.9**	**70.8**	**65.0**	**72.8**

Table 3: Detailed ImageQA and VideoQA performance of MLLMs on WaymoQA. **Bold**: best; underline: second best. Abbreviations: SC (Scene Caption), SU (Scene Understanding), PE (Perception), PR (Prediction), PL (Planning), TG (Temporal Grounding), UN (Uncertainty), FT (Fine-tuning).

with no separate per-item payment.

As summarized in Table 2, each scene is annotated by at least three annotators, with one annotation manager flagging guideline violations. Scenes with Fleiss' $\kappa < 0.7$ or ambiguous answers are re-annotated, resulting in agreement above 0.90 across all scene types in both Image QA and Video QA. Three independent verifiers further check ambiguity, excessive difficulty, and multiple plausible answers, reducing the verification flag rate from 34% to 0% after revision.

4 Experiments

4.1 Experimental Settings

We assess multiple families and sizes of state-of-the-art MLLMs and expert models, including variants of LLaVA (Liu et al., 2024b; Li et al., 2024a), InternVL (Wang et al., 2025; Zhu et al., 2025; Gao et al., 2024), Qwen-VL (Bai et al., 2025), Phi3.5-vision (Abdin et al., 2024), Gemma3 (Team et al., 2025), and OpenDriveVLA (Zhou et al., 2025b).

For **Image QA**, each example consists of eight synchronized camera views covering the front, side, and rear directions. For **Video QA**, each video clip spans 14 seconds, we uniformly sample 50 frames per video for evaluation.

4.2 Results of Image QA and Video QA

Table 3 reports the main benchmark results. Across pretrained models, safety-critical performance is consistently lower than normal-scene performance, showing that safety-critical reasoning requires more than generic scene recognition: models must interpret agent behavior, feasible actions, and downstream consequences.

Among pretrained models, Qwen2.5-VL performs best in both modalities. Qwen2.5-VL(32B) achieves the best pretrained overall accuracy, followed by Qwen2.5-VL(7B). However, most pretrained models remain weak on safety-critical questions, especially in VideoQA, suggesting that temporal context alone does not ensure reliable risk-aware reasoning.

Fine-tuning on WaymoQA substantially improves performance. Qwen2.5-VL(7B) improves from 62.4% to 72.9% on ImageQA and from 63.8% to 72.8% on VideoQA, with large gains on safety-critical questions. It also surpasses GPT-5.2 under the same MCQ protocol, suggesting that targeted WaymoQA supervision can be more effective than general-purpose model scale alone. Additional fine-tuning results on LLaVA-OneVision(7B) and InternVL3.5(8B) show consistent gains, indicating that WaymoQA provides transferable supervision across MLLM families.

4.3 Open-ended Evaluation

MCQ accuracy is our primary protocol because it provides objective fixed-answer evaluation without relying on LLM-as-a-judge scoring. This is especially important for safety-critical questions, where open-ended automatic metrics can obscure differences among weak pretrained models. We therefore use open-ended evaluation only as a complementary analysis of free-form reasoning.

As shown in Table 4, WaymoQA fine-tuning substantially improves GPT-based scoring, BLEU, ROUGE, and BERTScore, achieving the best overall and Safety-Critical scores across most metrics. However, pretrained-model trends are less clear than in the MCQ benchmark, supporting our use of human-written MCQ as the main evaluation protocol. These results indicate that WaymoQA supervision improves not only answer selection but also free-form safety-critical reasoning.

Model	Overall				Safety-Critical			
	GPT	B-4	R-L	BERT	GPT	B-4	R-L	BERT
LLaVA-OneVision(7B)	25.2	1.43	10.6	85.1	25.5	0.86	10.5	85.2
InternVL3.5(8B)	25.3	1.59	11.5	84.6	20.0	1.32	11.4	84.8
Qwen2.5-VL(7B)	25.5	1.60	11.3	84.7	23.0	1.12	10.2	84.7
Qwen2.5-VL(32B)	25.1	0.85	10.2	83.7	19.9	0.09	8.8	82.7
Gemma3(27B)	25.5	1.45	11.7	85.0	21.3	1.32	11.7	85.1
InternVL2-DA-DriveLM(4B)	23.1	2.96	15.7	85.9	17.5	2.70	15.8	85.8
Qwen2.5-VL(7B) (FT)	**40.4**	**7.25**	**29.0**	**88.1**	**34.0**	**4.72**	**23.1**	**87.1**

Table 4: Open-ended evaluation on WaymoQA. Abbreviations: B-4(BLEU-4), R-L(ROUGE-L).

4.4 Two-stage Consistency Evaluation

WaymoQA formulates safety-critical reasoning as a two-stage action-induced process, where the Stage 1 maneuver induces or exposes the Stage 2 risk (Sec. 3.3).

We evaluate this dependency on a safety-critical ImageQA subset containing 245 key frames with paired Stage 1 and Stage 2 questions. Both questions are anchored to the same key frame, identified by the same Waymo scenario token and frame index. Each Stage 2 question is paired with its corresponding Stage 1 question from the same key frame, such that Stage 2 reasoning is conditioned on the Stage 1 maneuver. The evaluation contains 394 paired instances, consisting of 244 prediction chains and 150 planning chains. VideoQA is excluded from this evaluation because equivalent paired two-stage chains were not constructed for the video subset.

To verify whether models truly follow this dependency, we introduce a stricter *joint correctness* evaluation. Specifically, a two-stage instance is counted as correct only if **both** Stage 1 and Stage 2 answers match the ground truth:

$$\mathrm{Acc}_{\text{2-stage}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[\hat{y}_i^{(1)} = y_i^{(1)} \wedge \hat{y}_i^{(2)} = y_i^{(2)}\right]. \quad (1)$$

This joint criterion captures the causal dependency of sequential risks; Stage 2 is evaluated together with the corresponding Stage 1 decision rather than as an independent question. As shown in Table 5, this joint evaluation causes a substantial drop for generic baselines (typically 7–14%), indicating difficulty in maintaining stage-to-stage consistency when the dependency is enforced. In contrast, the WaymoQA-fine-tuned Qwen2.5-VL-7B exhibits a markedly smaller degradation, suggesting that WaymoQA supervision improves *decision-conditioned* reasoning across immediate risk resolution (Stage 1) and induced secondary risk mitigation (Stage 2).

Model	Single-stage	Two-stage	Δ
GPT-5.2	67.8	60.9	-6.9
LLaVA-OneVision-7B	53.0	38.9	-14.1
Qwen2.5-VL-32B	65.5	54.6	-10.9
Gemma3-27B	63.1	49.5	-13.6
OpenDriveVLA	59.3	52.1	-7.2
Qwen2.5-VL-7B (FT)	**72.8**	**70.4**	**-2.4**

Table 5: Two-stage consistency evaluation.

4.5 Controlled Supervision Analysis

An improvement after fine-tuning on WaymoQA could in principle arise simply from exposure to additional driving-domain QA data, rather than from safety-critical supervision itself. We therefore conduct a matched-control experiment using the same Qwen2.5-VL-7B model, LoRA configuration, training budget, and evaluation protocol, while varying only the training data.

As shown in Table 6, removing safety-critical examples from WaymoQA decreases accuracy by 4.6 points overall and 5.6 points on safety-critical questions. A size-matched NuScenes-QA control performs substantially worse, trailing full WaymoQA training by 11.9 and 12.9 points, respectively. These results indicate that the observed improvement cannot be explained solely by additional driving-domain supervision and that safety-critical examples provide a specific benefit for risk-focused reasoning.

Training Data	Overall	Safety-Critical
WaymoQA (Full)	**67.8**	**65.8**
WaymoQA (Normal-only)	63.2	60.2
NuScenes-QA (Matched)	55.9	52.9
Δ Full – Normal-only	+4.6	+5.6
Δ WaymoQA – NuScenes-QA	+11.9	+12.9

Table 6: Matched-control analysis using the same Qwen2.5-VL-7B model and LoRA training setup.

4.6 Transfer and Generalization

We further examine whether WaymoQA supervision transfers beyond the main QA benchmark. First, we evaluate planning performance on the Waymo End-to-End validation split. As shown in Table 7, WaymoQA supervision improves ADE and RFS even without trajectory supervision, suggesting that safety-critical reasoning provides useful high-level decision signals for downstream driving behavior. Combining WaymoQA with trajectory supervision yields the best performance, indicating that language-based safety reasoning and geometric motion supervision are complementary.

Second, we evaluate cross-dataset transfer on DriveLM-nuScenes and NuPlanQA. As shown in Table 7, Qwen2.5-VL(7B) fine-tuned on WaymoQA consistently improves over the base model on external driving QA datasets. This suggests that the learned reasoning patterns are not narrowly overfitted to the WaymoQA distribution, but transfer to other autonomous-driving QA settings.

Finally, we conduct a preliminary closed-loop evaluation on Bench2Drive to examine whether WaymoQA supervision transfers beyond offline QA and open-loop planning. Using the same trajectory-supervised policy, WaymoQA improves the Driving Score from 65.0 to 71.2 and the Success Rate from 8.3% to 16.7% on 12 randomly selected routes. On 16 safety-critical routes, the Driving Score improves from 57.4 to 63.4 and the Success Rate from 6.2% to 12.5%. These results provide preliminary evidence that WaymoQA supervision transfers to downstream closed-loop behavior; however, given the limited number of evaluated routes, we treat them as supporting evidence rather than proof of deployable driving safety.

WaymoQA	Trajectory	ADE (5s) ↓	RFS ↑
✗	✗	14.30	5.37
✓	✗	11.60	5.91
✗	✓	4.23	6.37
✓	✓	**3.81**	**6.91**

Model	DriveLM MCQ	DriveLM GPT	NuPlan MCQ
Qwen2.5-VL(7B)	47.1	16.0	41.8
Qwen2.5-VL(7B) (FT)	**58.8**	**26.7**	**54.8**

Bench2Drive Setting	Base+Traj	+WaymoQA+Traj
Random (12 routes)	65.0 / 8.3	**71.2 / 16.7**
Safety-critical (16 routes)	57.4 / 6.2	**63.4 / 12.5**

Table 7: Transfer and generalization beyond the main WaymoQA benchmark. Bench2Drive results report Driving Score / Success Rate (%).

5 Analysis

5.1 Design Ablations

We analyze WaymoQA's core design choices, focusing on view coverage and temporal sampling.

Multi-view input. We compare single front view, front-3 views, and full 8-view inputs. As shown in Table 8, performance consistently improves as view coverage increases, especially when moving from front-only views to full surrounding views. This supports our motivation that safety-critical reasoning often depends on side and rear evidence that is unavailable from the front camera alone.

Temporal sampling rate. We compare our default 50-frame VideoQA setting, corresponding to approximately 1.2 FPS, with a denser 10 FPS setting. Table 8 shows only marginal differences across representative models, suggesting that 1.2 FPS is sufficient for the high-level reasoning targeted by WaymoQA. We emphasize that WaymoQA is a reasoning benchmark, not a proxy for real-time low-level control.

Model	**View Ablation** (ImageQA/VideoQA)			**FPS Ablation**	
	Front	Front-3	Full-8	1.2 FPS	10 FPS
LLaVA-OneVision(7B)	54.2 / 55.1	56.5 / 56.7	58.6 / 59.2	59.2	59.1
InternVL3.5(8B)	35.7 / 37.2	36.8 / 38.9	37.0 / 41.4	41.4	42.1
Qwen2.5-VL(7B)	58.1 / 56.2	58.8 / 59.3	62.4 / 63.8	63.8	64.0
Qwen2.5-VL(32B)	62.3 / 61.4	63.7 / 62.4	65.4 / 64.1	64.1	64.4
Gemma3(27B)	58.3 / 56.2	60.1 / 58.9	62.3 / 60.7	60.7	61.1
Qwen2.5-VL(7B) (FT)	**70.8 / 71.3**	**71.1 / 72.1**	**72.9 / 72.8**	**72.8**	**72.9**

Table 8: View and temporal sampling ablations.

5.2 Evaluation Robustness and Artifact Controls

To test answer-position bias, we evaluate each question under 10 random option permutations. Table 9 shows only small variance across models and modalities.

Question-only baseline. Qwen2.5-VL-7B achieves 39.7% without visual input and 40.5% after WaymoQA fine-tuning, whereas multimodal accuracy improves from 62.4% to 72.9% on ImageQA and from 63.8% to 72.8% on VideoQA. This suggests that the gains primarily rely on visual grounding rather than textual priors.

Blind surface-cue analysis. Simple heuristics based on question-option overlap, alphabetical order, punctuation, digit count, or fixed option selection achieve only 23.4–26.0%, close to the 25% chance level.

Distractor-swapped evaluation. On a held-out subset with regenerated distractors, the fine-tuning gain is preserved: overall accuracy improves from

62.7% to 70.6% with original distractors and from 75.0% to 83.4% with regenerated distractors. Thus, the improvement is not specific to the original distractor construction.

Together with the open-ended evaluation in Sec. 4.3, these results support the reliability of our MCQ protocol across answer permutations, textual-prior controls, distractor variations, and free-form generation.

Model	ImageQA		VideoQA	
	Mean	Var.	Mean	Var.
LLaVA-OneVision(7B)	58.4	0.13	58.9	0.07
InternVL3.5(8B)	37.0	0.02	41.1	0.03
Qwen2.5-VL(7B)	62.4	0.09	63.4	0.04
Qwen2.5-VL(32B)	65.2	0.06	64.1	0.02
Gemma3(27B)	62.1	0.04	60.9	0.06
InternVL2-DA-DriveLM(4B)	43.4	0.11	50.1	0.02
Qwen2.5-VL(7B) (FT)	**72.9**	0.09	**72.7**	0.03

Table 9: Robustness to answer-option ordering. We report mean and variance over 10 random permutations.

5.3 Environmental and Statistical Reliability

We analyze whether WaymoQA exhibits environmental imbalance and whether model performance varies across conditions. The dataset naturally reflects common real-world conditions, with more examples from bright daytime, sunny or overcast weather, light traffic, high visibility, and dry roads. Rarer conditions such as heavy rain, dark night, heavy traffic, and low visibility are less represented.

We further evaluate model performance across time of day, weather, road type, traffic density, visibility, and road condition. As shown in Table 10, performance varies across conditions, confirming that environment-aware analysis is important. At the same time, the relative ranking of models is broadly stable, and the WaymoQA fine-tuned model performs best across most conditions.

Attribute	QA %	Attribute	QA %	Attribute	QA %
Bright day	57.6	Sunny	52.7	Urban residential	38.2
Overcast day	20.0	Overcast	27.4	Intersection	30.3
Lit night	17.1	Partly cloudy	7.8	Urban downtown	17.9
Dusk/dawn	3.9	Light rain	5.6	Suburban	8.5
Dark night	1.5	Clear	3.7		
Light traffic	78.4	High visibility	86.2	Dry road	90.4
Moderate traffic	17.4	Medium visibility	13.4	Wet road	9.6
Empty traffic	4.0	Low visibility	0.4		
Heavy traffic	0.1				

Table 10: Environment distribution of WaymoQA.

To further test robustness, we conduct an augmentation-based evaluation on a 1K test-QA subset by generating environmental variants of the same scenes, including dusk/dawn, foggy, nighttime, overcast, rainy, and snowy conditions. Table 11 shows that performance degrades under adverse conditions, with different perturbations affecting Normal Scene and Safety-Critical questions differently. This suggests that environmental robustness remains an important future direction.

Finally, because some fine-grained subcategories contain relatively few samples, we conduct repeated evaluations on small subcategories and report the mean and variance in Table 12. The observed variances are small, suggesting that the main performance gaps are stable rather than artifacts of random fluctuations from small sample sizes.

Model	D/D	Fog	Night	Over	Rain	Snow
LLaVA-OV	-3.6/-2.4	-6.6/-1.7	-3.6/-2.0	-1.8/-6.5	-4.1/-6.8	-1.8/-7.5
Qwen-7B	-5.7/-2.7	-7.7/-0.0	-4.1/-3.5	-4.1/-3.5	-5.2/-5.9	-3.1/-4.6
Qwen-32B	-8.2/-0.3	-9.1/-2.4	-4.1/-0.0	-4.3/-0.6	-4.3/-7.3	-3.4/-5.9
Gemma3	-7.1/-4.0	-8.7/-3.1	-3.9/-1.3	-4.1/-2.7	-3.9/-5.1	-3.7/-2.4
Qwen-FT	-5.5/-7.5	-6.7/-7.8	-4.3/-5.6	-9.2/-6.2	-4.0/-3.9	-5.9/-7.3
Avg.	-6.02/-3.38	-7.76/-3.00	-4.00/-2.48	-4.70/-3.90	-4.30/-5.80	-3.58/-5.54

Table 11: Accuracy drop under environment augmentation on a 1K test-QA subset. Each cell reports Δ Normal / Δ Safety-Critical. Abbreviations: D/D (Dusk/Dawn), Fog (Foggy), Over (Overcast).

Cat.	Sub.	LLaVA1.5	InternVL3	InternVL3.5	Qwen FT
Safety-Critical	SC	13.8 / 0.6	40.5 / 0.0	29.7 / 0.0	63.5 / 0.0
Safety-Critical	PE	32.6 / 0.0	50.0 / 0.0	52.1 / 0.31	63.5 / 0.0
Safety-Critical	TG	27.4 / 0.0	25.4 / 0.0	19.6 / 0.0	58.9 / 0.0
Safety-Critical	UN	33.3 / 0.0	56.2 / 0.0	43.7 / 0.0	71.0 / 0.39

Table 12: Repeated-run reliability on small VideoQA sub-categories. We report the mean and variance over 10 runs. Abbreviations: SC (Scene Caption), PE (Perception), TG (Temporal Grounding), UN (Uncertainty).

6 Conclusion

WaymoQA reveals a clear gap between general vision-language ability and safety-critical driving reasoning. Pretrained MLLMs struggle when safety-critical scenes require connecting perception, prediction, and planning under action-induced downstream risks, highlighting the importance of reasoning about the consequences of ego actions.

Fine-tuning on WaymoQA consistently improves ImageQA, VideoQA, open-ended reasoning, and two-stage consistency, while multi-view ablations confirm the importance of surrounding context for safety-critical decisions. However, performance degradation under adverse conditions indicates that robust safety-critical reasoning remains an open challenge.

Overall, WaymoQA provides a 35K human-annotated multi-view benchmark and training resource that formulates safety-critical reasoning as immediate-risk resolution followed by downstream-risk mitigation.

Limitations

WaymoQA has several limitations. **First**, WaymoQA is built from a single source, the Waymo End-to-End dataset, and therefore reflects its geographic, sensor, and scenario distributions. Here, *long-tail* refers to rare NHTSA pre-crash scenario types rather than rare environmental conditions. As shown in our environmental analysis, common conditions such as bright daytime, dry roads, and light traffic are more frequent, while heavy traffic, low visibility, dark night, and adverse weather remain underrepresented. This limits the diversity of safety-critical driving conditions covered by the benchmark.

Second, WaymoQA primarily evaluates high-level safety-critical reasoning rather than closed-loop driving control. Although our planning-transfer and preliminary Bench2Drive results suggest that WaymoQA supervision can benefit downstream driving behavior, the closed-loop evaluation is limited in scale. Larger-scale simulation and real-world validation are therefore needed before drawing conclusions about deployable driving safety.

Third, our main evaluation relies on human-written multiple-choice questions. This design enables objective and reproducible scoring, but it may simplify the space of valid driving decisions and cannot fully capture open-ended safety-critical reasoning. We therefore include open-ended evaluation as a complementary analysis, although automatic free-form metrics remain imperfect for assessing causal reasoning and action justification.

Finally, WaymoQA decomposes safety-critical reasoning into structured question categories. While this enables diagnosis of failures in perception, prediction, planning, and uncertainty, it does not fully enforce a single end-to-end reasoning chain from visual evidence to final action. Future work can extend WaymoQA toward compound multi-step reasoning with partial-credit evaluation.

Ethical Considerations

WaymoQA is built from the Waymo End-to-End dataset and follows the original Waymo data license and access terms. We will release annotations and code for research use, while raw images and videos must be accessed through the original Waymo dataset.

The source Waymo E2E data are distributed under the privacy and access procedures of the original Waymo dataset. We do not redistribute raw images or videos, and users must obtain them through the original dataset release. Our released annotations consist of question-answer pairs, category labels, and evaluation metadata, and do not include personally identifying textual information. During annotation and verification, we also check that QA contents do not name or uniquely identify individual people and do not contain offensive or irrelevant content.

Acknowledgments

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Korea government (MSIT) (No. RS-2025-00520207); Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (Nos. RS-2024-00457882, 2022-0-01045, 2022-0-00680, RS-2019-II190075); the Advanced GPU Utilization Support Program funded by the Government of the Republic of Korea (Ministry of Science and ICT); the AI Computing Infrastructure Enhancement (GPU Rental Support) User Support Program funded by the Ministry of Science and ICT (MSIT), Republic of Korea (No. RQT-25-120217); a grant partly supported by both IITP (MSIT) and Korea Evaluation Institute of Industrial Technology (KEIT) (MOTIE) (No. RS-2025-02217259); and the next-generation CultureTech technology development (R&D) project for cultural space AX transformation through the Korea Creative Content Agency (KOCCA) grant funded by the Ministry of Culture, Sports and Tourism (MCST) in 2026 (No. RS-2026-25508598).

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

National Highway Traffic Safety Administration and 1 others. 2007. Pre-crash scenario typology for crash avoidance research. *DOT HS*, 810:767.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie

Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024a. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024b. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.

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

Wei-Lin Chen, Zhepei Wei, Xinyu Zhu, Shi Feng, and Yu Meng. Do llm evaluators prefer themselves for a reason?, 2025. *URL https://arxiv.org/abs/2504.03846*, 2504.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. 2024. Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 902–909.

Xinpeng Ding, Jianhua Han, Hang Xu, Wei Zhang, and Xiaomeng Li. 2025. Hilm-d: Enhancing mllms with multi-scale high-resolution details for autonomous driving. *International Journal of Computer Vision*, pages 1–17.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR.

Jianwu Fang, Lei-lei Li, Junfei Zhou, Junbin Xiao, Hongkai Yu, Chen Lv, Jianru Xue, and Tat-Seng Chua. 2024. Abductive ego-view accident video understanding for safe driving perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22030–22040.

Bowen Feng, Zhiting Mei, Baiang Li, Julian Ost, Roger Girgis, Anirudha Majumdar, and Felix Heide. 2025. Verdi: Vlm-embedded reasoning for autonomous driving. *arXiv preprint arXiv:2505.15925*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118.

Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. 2024. Drive like a human: Rethinking autonomous driving with large language models. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 910–919. IEEE.

Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, and 1 others. 2024. Mini-internvl: A flexible-transfer pocket multimodal model with 5% parameters and 90% performance. *arXiv preprint arXiv:2410.16261*.

Xianda Guo, Ruijun Zhang, Yiqun Duan, Yuhang He, Chenming Zhang, Shuai Liu, and Long Chen. 2024. Drivemllm: A benchmark for spatial understanding with multimodal large language models in autonomous driving. *arXiv e-prints*, pages arXiv–2411.

Ziang Guo, Konstantin Gubernatorov, Selamawit Asfaw, Zakhar Yagudin, and Dzmitry Tsetserukou. 2025. Vdt-auto: End-to-end autonomous driving with vlm-guided diffusion transformers. *arXiv preprint arXiv:2502.20108*.

Wencheng Han, Dongqian Guo, Cheng-Zhong Xu, and Jianbing Shen. 2025. Dme-driver: Integrating human decision logic and 3d scene perception in autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3347–3355.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, and 1 others. 2024. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14281–14290.

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

Zhijian Huang, Tao Tang, Shaoxiang Chen, Sihao Lin, Zequn Jie, Lin Ma, Guangrun Wang, and Xiaodan Liang. 2024. Making large language models better planners with reasoning-decision alignment. In *European Conference on Computer Vision*, pages 73–90. Springer.

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

Keishi Ishihara, Kento Sasaki, Tsubasa Takahashi, Daiki Shiono, and Yu Yamaguchi. 2026. Stride-qa: visual question answering dataset for spatiotemporal reasoning in urban driving scenes. In *Proceedings of*

the AAAI Conference on Artificial Intelligence, volume 40, pages 5257–5266.

Ye Jin, Ruoxuan Yang, Zhijie Yi, Xiaoxi Shen, Huiling Peng, Xiaoan Liu, Jingli Qin, Jiayang Li, Jintao Xie, Peizhong Gao, and 1 others. 2024. Surrealdriver: Designing llm-powered generative driver agent framework based on human drivers' driving-thinking data. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 966–971. IEEE.

Boshra Khalili and Andrew W Smyth. 2025. Autodrive-qa-automated generation of multiple-choice questions for autonomous driving datasets using large vision-language models. *arXiv preprint arXiv:2503.15778*.

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

Younggun Kim, Ahmed S Abdelrahman, and Mohamed Abdel-Aty. 2025. Vru-accident: A vision-language benchmark for video question answering and dense captioning for accident scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 761–771.

Yuwon Lee. 2024. Qwen2-vl-finetune.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Cheng Li, Keyuan Zhou, Tong Liu, Yu Wang, Mingqiao Zhuang, Huan-ang Gao, Bu Jin, and Hao Zhao. 2025a. Avd2: Accident video diffusion for accident video description. *arXiv preprint arXiv:2502.14801*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Yue Li, Meng Tian, Zhenyu Lin, Jiangtong Zhu, Dechang Zhu, Haiqiang Liu, Yueyi Zhang, Zhiwei Xiong, and Xinhai Zhao. 2025b. Fine-grained evaluation of large vision-language models in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9431–9442.

Haohong Lin, Xin Huang, Tung Phan, David Hayden, Huan Zhang, Ding Zhao, Siddhartha Srinivasa, Eric Wolff, and Hongge Chen. 2025. Causal composition diffusion model for closed-loop traffic generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27542–27552.

Haolan Liu, Liangjun Zhang, Siva Kumar Sastry Hari, and Jishen Zhao. 2024a. Safety-critical scenario generation via reinforcement learning based editing. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14405–14412. IEEE.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024c. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Zuyan Liu, Yuhao Dong, Yongming Rao, Jie Zhou, and Jiwen Lu. 2024d. Chain-of-spot: Interactive reasoning improves large vision-language models. *arXiv preprint arXiv:2403.12966*.

Haohan Luo and Feng Wang. 2023. A simulation-based framework for urban traffic accident detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. 2024. Dolphins: Multimodal language model for driving. In *European Conference on Computer Vision*, pages 403–420. Springer.

Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. 2023. Drama: Joint risk localization and captioning in driving. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1043–1052.

Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. 2023. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*.

Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, and 1 others. 2024. Lingoqa: Visual question answering for autonomous driving. In *European Conference on Computer Vision*, pages 252–269. Springer.

Chirag Parikh, Deepti Rawat, Tathagata Ghosh, Ravi Kiran Sarvadevabhatla, and 1 others. 2025. Roadsocial: A diverse videoqa dataset and benchmark for road event understanding from social video narratives. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19002–19011.

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. *arXiv preprint arXiv:2503.12772*.

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

Davis Rempe, Jonah Philion, Leonidas J Guibas, Sanja Fidler, and Or Litany. 2022. Generating useful accident-prone driving scenarios via a learned traffic prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17305–17315.

Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. 2024a. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15120–15130.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024b. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Chonghao Sima, Kashyap Chitta, Zhiding Yu, Shiyi Lan, Ping Luo, Andreas Geiger, Hongyang Li, and Jose M Alvarez. 2025. Centaur: Robust end-to-end autonomous driving with test-time training. *arXiv preprint arXiv:2503.11650*.

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.

Austin Stone, Ted Xiao, Yao Lu, Keerthana Gopalakrishnan, Kuang-Huei Lee, Quan Vuong, Paul Wohlhart, Sean Kirmani, Brianna Zitkovich, Fei Xia, and 1 others. 2023. Open-world object manipulation using pre-trained vision-language models. *arXiv preprint arXiv:2303.00905*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. 2024. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*.

Tianqi Wang, Sukmin Kim, Ji Wenxuan, Enze Xie, Chongjian Ge, Junsong Chen, Zhenguo Li, and Ping Luo. 2024. Deepaccident: A motion and accident prediction benchmark for v2x autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5599–5606.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*.

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. *arXiv preprint arXiv:2501.04003*.

Yuting Xie, Xianda Guo, Cong Wang, Kunhua Liu, and Long Chen. Advdiffuser: Generating adversarial safety-critical driving scenarios via guided diffusion. in 2024 ieee. In *RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9983–9989.

Runsheng Xu, Hubert Lin, Wonseok Jeon, Hao Feng, Yuliang Zou, Liting Sun, John Gorman, Kate Tolstaya, Sarah Tang, Brandyn White, and 1 others. 2025. Wod-e2e: Waymo open dataset for end-to-end driving in challenging long-tail scenarios. *arXiv preprint arXiv:2510.26125*.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. 2024. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*.

Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Yuchen Wang, Ella Atkins, and David J Crandall. 2022. Dota: Unsupervised detection of traffic anomaly in driving videos. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):444–459.

Tong Zeng, Longfeng Wu, Liang Shi, Dawei Zhou, and Feng Guo. 2025. Are vision llms road-ready? a comprehensive benchmark for safety-critical driving video understanding. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5972–5983.

Jiawei Zhang, Chejian Xu, and Bo Li. 2024. Chatscene: Knowledge-enabled safety-critical scenario generation for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15459–15469.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou,

Tianyi Li, and Yang You. 2024. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*.

Ziyuan Zhong, Davis Rempe, Yuxiao Chen, Boris Ivanovic, Yulong Cao, Danfei Xu, Marco Pavone, and Baishakhi Ray. 2023. Language-guided traffic simulation via scene-level diffusion. In *Conference on robot learning*, pages 144–177. PMLR.

Ziyuan Zhong, Davis Rempe, Danfei Xu, Yuxiao Chen, Sushant Veer, Tong Che, Baishakhi Ray, and Marco Pavone. 2022. Guided conditional diffusion for controllable traffic simulation. *arXiv preprint arXiv:2210.17366*.

Jiawei Zhou, Linye Lyu, Zhuotao Tian, Cheng Zhuo, and Yu Li. 2025a. Safemvdrive: Multi-view safety-critical driving video synthesis in the real world domain. *arXiv preprint arXiv:2505.17727*.

Xingcheng Zhou, Xuyuan Han, Feng Yang, Yunpu Ma, Volker Tresp, and Alois Knoll. 2025b. Opendrivevla: Towards end-to-end autonomous driving with large vision language action model. *arXiv preprint arXiv:2503.23463*.

Zewei Zhou, Tianhui Cai, Seth Z Zhao, Yun Zhang, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. 2025c. Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. *arXiv preprint arXiv:2506.13757*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Appendix Overview

This appendix provides the following supplementary materials:

- **Implementation details** (Appendix A): LoRA setup, training hyperparameters, sampling strategy, and hardware configuration.
- **Dataset documentation** (Appendix B): question-category statistics, task definitions, and representative templates.
- **MCQ construction** (Appendix C): criteria for converting human annotations into four-option multiple-choice questions.
- **Complete benchmark tables** (Appendix D): full ImageQA and VideoQA results for all models and categories.
- **Additional ablations** (Appendix E): training-modality ablations comparing Image-only, Video-only, and joint Image&Video fine-tuning.
- **Open-ended evaluation** (Appendix F): implementation details, prompts, package settings, and full GPTScore, BLEU, ROUGE, and BERTScore results across scene groups.
- **Environmental augmentation examples** (Appendix G): qualitative examples of appearance changes under dusk/dawn, foggy, nighttime, overcast, rainy, and snowy conditions.
- **Failure analysis** (Appendix H): a representative two-stage reasoning failure involving relative spatial directions.
- **Extending WaymoQA** (Appendix I): future extensions toward compound multi-step reasoning and partial-credit evaluation.

A Fine-Tuning and Implementation Details

A.1 Fine-Tuning Details

We fine-tune Qwen2.5 VL(7B) following the official guideline (Lee, 2024). The base weights of the language and vision towers and the projector are frozen, and LoRA adapters are applied to all supported modules except the LM head and token embeddings. The batch size is 32 and learning rate is 2e-4. For data construction, videos are packed with all QA pairs associated with the clip and key frames are packed with all QA pairs anchored to the frame.

For training stability, we sample video at 1.2 FPS and train with a multi image fine tuning setup analogous to Image QA. Video data consist of 22 second multi view clips that impose substantial memory demand; accordingly, we use 1.2 FPS under the available hardware constraints. We expect non-uniform sampling to improve learning by allocating samples to key frames and fewer to less informative frames. In practice the 28,585 training pool benefited from allocating at least 20,000 examples to training. Validation sets larger than 8,000 were associated with lower accuracy.

A.2 Implementation Details

We fine-tune Qwen2.5-VL-7B-Instruct on WaymoQA using 8×NVIDIA RTX A6000 GPUs with DeepSpeed ZeRO-2. We adopt parameter-efficient fine-tuning with LoRA following the official guideline (Lee, 2024). Both language and vision modules are equipped with LoRA adapters (rank 64, $\alpha = 64$, dropout 0.05), while the base vision tower, language model, and merger module are frozen. Following the official Qwen2.5-VL setup, we exclude the token embedding and output head from LoRA (`lm_head`, `embed_tokens`) and keep all other linear layers trainable through adapters. We enable Liger kernels for efficient attention, gradient checkpointing for memory savings, and train in `bfloat16` with `tf32` enabled on matrix multiplications.

Training uses a global batch size of 32 (per-device batch size 1 on 8 GPUs with gradient accumulation of 4), for 3 epochs over the WaymoQA training split. We apply a cosine learning rate schedule with an initial learning rate of 2×10^{-4}, a warm-up ratio of 0.03, and weight decay of 0.1. Images are dynamically resized within the range of $256 \times 28 \times 28$ to $1024 \times 28 \times 28$ pixels to balance resolution and memory usage.

B Question Categories and Statistics

Table 13 and Figure 5 summarizes how the 62-item question bank is instantiated across **Format** (Image / Video), **Scene type** (Normal / Safety-critical / Others), and **Task** categories.

Task definitions. Below we briefly describe each reasoning category covered by the question bank.

Caption. These questions ask the model to

Type	Scene	Category	Count
Image	Normal	Normal Perception & Prediction	4,004
Image	Normal	Normal Planning	1,518
Image	Normal	Normal scene Understanding	4,606
Image	Normal	Normal scene caption	1,134
Image	Normal	Traffic Signal	1,266
Image	Others	Counterfactual	1,616
Image	Others	Spatial Relationship	351
Image	Others	Special Event	1,927
Image	Safety-critical	Safety-critical Perception	1,375
Image	Safety-critical	Safety-critical Planning	3,239
Image	Safety-critical	Safety-critical Prediction	2,990
Image	Safety-critical	Safety-critical scene Understanding	1,926
Image	Safety-critical	Safety-critical scene caption	1,060
Image	Safety-critical	Uncertainty	2,317
Video	Normal	Normal Perception & Prediction	1,080
Video	Normal	Normal scene Understanding	1,270
Video	Normal	Normal scene caption	290
Video	Normal	Traffic Signal	300
Video	Others	Counterfactual	316
Video	Others	Spatial Relationship	143
Video	Others	Special Event	202
Video	Safety-critical	Safety-critical Perception	324
Video	Safety-critical	Safety-critical scene Understanding	874
Video	Safety-critical	Safety-critical scene caption	290
Video	Safety-critical	Temporal Grounding	316
Video	Safety-critical	Uncertainty	288
Sum			35,022

Table 13: Distribution of questions in WaymoQA by format, scene type, and task category.

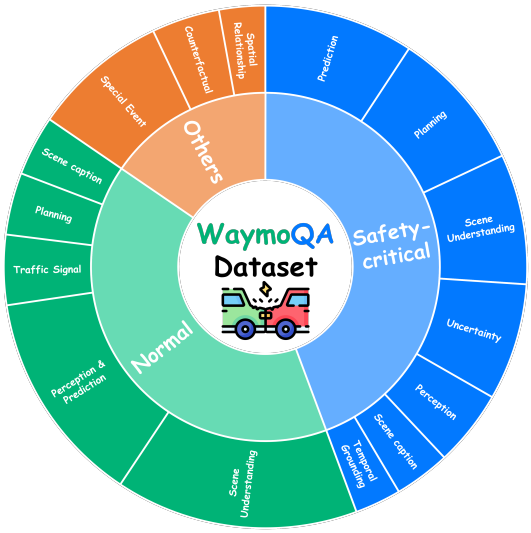


Figure 5: VQA distribution in the WaymoQA dataset.

produce or select a caption summarizing the scene. For *normal* scenes, the caption focuses on overall driving context such as road layout, traffic participants, and general situation. For *safety critical* scenes, the caption is required to highlight the safety relevant situation, such as an imminent collision risk or a blocked lane.

Scene Understanding. Compared to Caption, these questions probe more fine grained comprehension of the scene. They query specific aspects of the environment such as which lanes are drivable, who has right of way, and where the ego vehicle is positioned rather than a single high level summary.

Perception. Perception questions ask about properties or states of objects in the scene: their type such as car, pedestrian, cyclist, or construction vehicle, their attributes such as parked or moving or indicating, and their positions relative to the ego vehicle or road structure.

Prediction. Prediction questions require anticipating the *future behavior* of agents. Typical queries include whether another vehicle will cut in, whether a pedestrian is likely to cross, or how surrounding traffic will evolve in the next few seconds.

Planning. Planning questions focus on the ego vehicle's *future action*. The model must decide an appropriate maneuver such as continue, slow down, stop, or change lanes, given the perceived scene and expected motions of other agents.

Uncertainty. These questions target reasoning about unknown or unobserved risk, such as occluded areas, blind spots, or partially visible agents. The model must recognize where it cannot be certain, for example when a pedestrian may emerge from behind a truck, and reflect this uncertainty in its choice.

Counterfactual. Counterfactual questions describe hypothetical deviations from the current situation. They typically ask what would happen *if* the ego vehicle or another agent took a different action, for example what would likely occur if the ego vehicle changed lanes now, probing causal reasoning about alternative behaviors.

Special Event. Special Event questions focus on unusual but practically important conditions, such as road construction, emergency vehicles, accidents, or adverse weather such as rain, snow, or poor visibility. The model must detect these events and understand their impact on safe driving.

Spatial Relationship. These questions query the interaction and relative positioning between two or more objects. Examples include determining which vehicle is ahead or behind, which lane another car occupies, or whether two agents are on a collision course.

Temporal Grounding. Temporal Grounding

Category	Representative question templates
Normal Prediction & Planning	Based on the current appearance of the person or two-wheeled vehicle, how do they intend to move? Which of the following describes a non-safety-critical object?
Normal Planning	What actions are currently available?
Normal Scene Understanding	Which of the following statements about traffic density is correct? How is the behavior of the vehicle summarized before the event?
Normal Scene Caption	Which of the following is correct as an image caption from the perspective of a normal frame? What is the video summary from the perspective of a normal scene?
Counterfactual	What situation might occur if the vehicle acted this way? What are the potential situations that could occur in the current scenario?
Spatial Relationship	What is the spatial relationship of two or more objects?
Special Event	Is there a special situation? What are the restrictions on behavior in school zones or congested areas? What is the plan for yielding when an ambulance or fire truck approaches?
Safety-critical Perception	What is a safety-critical object?
Safety-critical Planning	What are the driving guidelines for the ego vehicle based on the behavior of safety-critical objects? How should the ego vehicle drive to solve the first hazard while also addressing the second hazard?
Safety-critical Prediction	How do you think a safety-critical object will move? Are there any other safety-critical objects besides the first safety-critical object? If so, how would they move?
Safety-critical Scene Understanding	What is the correlation in safety-critical situations? How did the safety-critical object move and how did the ego vehicle behave?
Safety-critical Scene Caption	What is the correct image caption from the perspective of a safety-critical frame? What is correct regarding video summary from a safety-critical scene perspective?
Uncertainty	Where is the area that hides the danger in the current frame? Is there a blind spot in the field of view?
Temporal Grounding	Where is the safety-critical situation in the video?
Traffic Signal	Which of the following is a correct description of traffic signals?

Table 14: Representative question templates for each reasoning category in the WaymoQA question bank.

questions ask *when* in a video the situation becomes safety critical. The model must identify the time interval or frame range where risk emerges or peaks, rather than only describing a single static frame, and this category is defined for video data only.

Traffic Signal. These questions assess correct interpretation of traffic lights, arrows, and lane markings and how they constrain ego behavior. They often require relating signal state such as red light, protected left arrow, or lane only sign to valid or invalid driving actions.

We further provide representative question templates for each reasoning category in Table 14.

C Criteria of Converting Annotation into MCQ

In order to construct the multiple-choice questions (MCQ) used in WaymoQA, we convert human-annotated answers into a structured four-choice format. Each MCQ requires one correct answer and three plausible false candidates. The false answers must be carefully designed so that the overall difficulty remains appropriate. Our target difficulty range ensures that pretrained models score roughly between 30% and 70% because scores below 30% indicate that the question is unnecessarily difficult, and scores above 70% mean that the question is too easy. We iteratively refine the criteria by creating subsets of MCQs, and adjusting the false-choices construction rules. The final criteria are summarized below.

1. The question must not be solvable from text input alone. False choices that are irrelevant, exaggerated, or logically disconnected allow models to answer correctly without looking at the image. To avoid this issue, all candidates must remain close to the topic of the question. Each false answer should carry plausible content that resembles the correct answer so that visual understanding is required.

2. False choices must be visually similar to the

correct answer. To ensure that the MCQ reflects genuine visual reasoning, false candidates must be close variants of the correct answer. We modify only minor attributes such as object type, trajectory, color, count, or location. These changes create subtle perturbations that require precise image understanding. For example:

- **True**: *"A black vehicle is going straight."* → **False**: *"A white vehicle is turning right."*
- **True**: *"Two cars are parked on the right side of the road."* → **False**: *"Five cars are parked on the left side of the road."*

3. Each question must have exactly one correct answer. We manually verify that no false candidate can be interpreted as correct. If a candidate introduces ambiguity or can be considered a valid answer under certain contexts, we revise or discard it.

4. All choices must follow a consistent writing style. Since multiple annotators contribute to the dataset, differences in tone and phrasing naturally occur. A final reviewer rewrites all answer options to unify the style across the dataset. This process ensures that sentence structure, wording, and perspective remain consistent across all choices.

D Complete Benchmark Tables

The main paper reports a compact table that merges ImageQA and VideoQA. Here we provide the original detailed ImageQA and VideoQA tables before merging. Table 15 and 16 show all Normal, Others, Safety-Critical, and Overall category scores.

E Ablation Study

Ablation on training modality. Table 17 analyzes how the choice of training modality affects performance on WaymoQA. Starting from the off-the-shelf Qwen2.5-VL(7B) baseline, we fine-tune three variants: (i) on both Image and Video QA data, (ii) on Image QA only, and (iii) on Video QA only.

For **Image QA**, all fine-tuned models outperform the baseline (62.4% overall), but the joint Image&Video model is clearly the most effective. Training only on images raises overall Image QA accuracy to 65.7%, while training only on videos yields 63.4%. In contrast, using both modalities pushes Image QA to 72.9% overall and improves every scene type: normal scenes from 68.1% to 74.9%, "Others" from 55.6% to 72.7%, and safety-critical scenes from 61.1% to 71.2%. Thus, even for image-only evaluation, exposure to video questions provides a substantial gain, especially on safety-critical categories.

A similar trend appears for **Video QA**. Fine-tuning on video-only data improves the baseline from 62.6% to 65.1% overall, and image-only training gives a smaller boost to 63.9%. However, the Image&Video model again achieves the best result with 72.8% overall accuracy, lifting normal scenes from 72.7% to 76.0%, "Others" from 65.2% to 83.1%, and safety-critical scenes from 51.3% to 65.0%.

These results confirm our hypothesis from the main text: while modality-specific fine-tuning helps the corresponding evaluation modality, jointly training on both image and video questions yields the largest improvements, particularly for safety-critical reasoning, by encouraging the model to share representations across complementary temporal and spatial cues.

F Open-ended Evaluation

In the main paper, we report a compact open-ended evaluation to complement the primary MCQ benchmark. Here, we provide the full open-ended results for each scene group: Normal Scene, Others, Safety-Critical Scene, and Overall. This analysis verifies whether the gains observed in the MCQ protocol also transfer to free-form answer generation.

We evaluate representative models using GPT-based scoring, BLEU-1/2/3/4, ROUGE-1/2/L, and BERTScore. Across all scene groups, the WaymoQA fine-tuned Qwen2.5-VL(7B) model achieves the strongest performance on GPTScore and most lexical/semantic metrics. This confirms that WaymoQA supervision improves not only answer-option selection, but also open-ended safety-critical reasoning. At the same time, the relative ordering among pretrained models is less consistent than in the MCQ benchmark, supporting our use of human-written MCQs as the primary evaluation protocol and open-ended scoring as complementary evidence.

Model	Normal Scene						Others			Safety - Critical							Overall
	TS	SC	SU	PL	PP	Overall	CF	SE	Overall	SU	PE	PR	PL	SC	UN	Overall	
Human Experts	96.6	97.7	97.7	95.3	94.0	96.1	94.8	96.7	96.3	96.8	98.0	96.4	94.3	98.7	94.9	96.0	96.1
GPT 5.2	82.0	88.9	73.1	71.7	75.4	75.7	56.4	57.1	57.0	68.4	65.6	73.8	62.3	57.6	53.0	64.5	67.4
LLaVA1.5(7B)	19.6	24.4	50.0	24.4	32.8	35.9	30.7	39.0	37.5	24.5	21.3	24.7	25.5	27.2	17.0	23.2	30.7
LLaVA-OneVision(7B)	71.3	78.5	**73.3**	65.5	56.1	66.7	53.6	51.4	51.7	50.8	42.2	56.9	53.2	45.5	**64.9**	54.4	58.6
InternVL3.5(8B)	14.7	31.8	45.3	60.8	37.4	40.1	41.3	47.9	46.6	26.6	30.4	23.7	41.1	25.3	22.2	28.9	37.0
InternVL3.5(4B)	17.6	32.5	42.5	43.0	39.6	38.0	40.2	42.8	42.3	21.7	24.1	23.6	39.2	41.1	20.2	27.8	34.4
InternVL3(8B)	16.8	29.6	43.4	46.8	24.8	33.6	34.0	49.9	47.0	28.8	24.9	27.1	46.2	41.7	27.8	32.9	35.9
InternVL3(14B)	20.2	31.2	45.9	50.0	39.2	40.2	42.1	58.6	55.6	23.4	32.8	52.1	45.3	45.8	39.0	42.1	44.0
Qwen2.5-VL(32B)	79.9	**81.1**	70.7	64.4	69.4	71.2	50.0	62.6	60.3	52.6	44.6	74.3	64.6	56.4	61.4	62.6	65.4
Qwen2.5-VL(7B)	77.4	77.4	71.2	48.4	63.3	68.1	43.5	58.9	55.6	54.4	55.3	71.0	64.0	43.0	54.6	61.1	62.4
Phi3.5-vision	24.1	32.5	63.2	61.2	40.6	48.9	44.6	44.5	44.5	32.0	25.6	41.8	45.6	44.3	46.5	40.9	44.7
Gemma3(4B)	60.2	62.6	62.6	56.2	57.1	59.7	48.0	48.9	48.7	34.2	37.1	53.8	44.3	51.3	33.8	43.5	50.6
Gemma3(27B)	67.6	73.3	68.7	51.5	61.5	64.3	49.1	65.6	62.6	67.2	61.6	67.1	59.4	54.4	49.4	60.4	62.3
InternVL2-DA-DriveLM(4B)	34.0	58.5	60.2	**66.2**	49.6	54.2	34.6	38.5	37.7	44.4	24.1	41.2	26.8	38.6	46.1	36.8	43.6
OpendriveVLA	82.0	76.0	42.1	50.3	57.1	54.9	41.3	44.5	43.9	57.1	50.7	68.3	53.1	41.1	49.6	56.0	53.4
Qwen2.5-VL(7B) (FT, Ours)	**84.4**	72.6	72.0	58.5	**80.6**	**74.9**	**63.1**	**74.8**	**72.7**	**76.1**	**76.6**	**77.8**	**65.9**	**67.8**	63.8	**71.2**	**72.9**

Table 15: Detailed Image QA performance of the MLLMs on WaymoQA. **Bold**: best; underline: second best. The following abbreviations are used: TS (Traffic Signal), SC (Scene Caption), SU (Scene Understanding), PL (Planning), PE (Perception), PR (Prediction), PP (Perception&Prediction), CF (Counterfactual), SE (Special Event), UN (Uncertainty), FT (Fine-tuning).

F.1 Evaluation Pipeline

We implement open-ended evaluation as a three-stage pipeline. First, the evaluated VLM generates a free-form answer from the visual input and question. Second, a GPT-based judge assigns a semantic correctness score to the generated answer with respect to the human-written reference answer. Third, we compute automatic text metrics in batch and aggregate the results by scene group.

For open-ended generation, the model receives only the visual input and the question. The multiple-choice answer options are not provided to the model. This prevents open-ended evaluation from becoming answer-option selection and instead tests whether the model can produce the answer in natural language. For examples originally stored in MCQ format, we recover the textual reference by selecting the ground-truth option and removing the option prefix such as "A." or "B.".

Input images are resized so that the maximum side length is at most 1024 pixels and encoded as JPEG with quality 85 before being passed to the model. The camera order follows a fixed priority order: FRONT, FRONT_LEFT, FRONT_RIGHT, REAR_LEFT, REAR_RIGHT, REAR, SIDE_LEFT, and SIDE_RIGHT. For generation, we use temperature 0.0, top-$p = 1.0$, and a maximum generation length of 64 tokens.

F.2 GPT-based Scoring

We use a GPT-based judge to assess semantic correctness beyond exact lexical overlap. The judge is instructed to act as an expert evaluator for autonomous-driving perception and planning. Each prediction is scored on a 0–100 scale using three criteria: factual correctness, completeness, and clarity/relevance. Factual correctness receives 50 points, completeness receives 30 points, and clarity/relevance receives 20 points. This weighting emphasizes whether the generated answer captures the correct driving fact, object state, or ego maneuver, which is especially important for safety-critical perception and planning questions.

The judge uses deterministic decoding with temperature 0.0 and a maximum output length of 4 tokens. It is required to return only a single integer between 0 and 100. For planning questions, correctness of the final maneuver is treated as the most important factor. Hallucinated objects or events are penalized, and verbose answers are not rewarded when the core driving fact is incorrect.

F.3 Automatic Text Metrics

We compute BLEU-1/2/3/4, ROUGE-1/2/L, and BERTScore for open-ended evaluation. BLEU is computed with NLTK sentence-level BLEU using smoothing method 1. ROUGE is computed with the `rouge-score` package, where we report ROUGE-1, ROUGE-2, and ROUGE-L F1

Model	Normal Scene					Others				Safety - Critical						Overall
	TS	SC	SU	PP	Overall	CF	SE	SR	Overall	SC	SU	PE	TG	UN	Overall	
Human Experts	95.4	95.7	100.0	98.7	96.9	95.9	100.0	90.5	96.6	97.2	97.2	96.0	94.0	92.5	95.9	96.5
GPT 5.2	74.0	70.6	76.3	85.3	79.1	82.0	89.4	47.6	78.8	54.1	65.5	50.0	49.0	81.3	61.6	72.6
LLaVA1.5(7B)	26.0	29.4	48.8	40.4	42.4	30.0	40.4	33.3	34.7	13.5	27.5	32.6	27.4	33.3	27.5	35.9
LLaVA-OneVision(7B)	80.0	35.2	68.8	60.7	65.8	**80.0**	93.6	**61.9**	82.1	32.4	43.4	40.3	29.4	62.5	42.3	59.2
InternVL3.5(8B)	30.0	17.6	48.8	47.8	45.1	52.0	61.7	23.8	50.8	29.7	28.2	51.9	19.6	43.7	32.9	41.4
InternVL3.5(4B)	26.0	41.1	44.1	28.8	36.3	62.0	63.8	38.0	58.4	37.8	23.4	42.3	17.6	39.5	29.3	36.7
InternVL3(8B)	26.0	35.2	38.1	37.4	36.3	58.0	29.7	19.0	39.7	40.5	27.5	50.0	25.4	56.2	36.2	36.8
InternVL3(14B)	28.5	50.0	51.2	46.0	46.6	72.0	60.4	33.3	60.4	59.4	27.3	51.9	25.0	57.4	38.6	46.3
Qwen2.5-VL(32B)	80.0	52.9	70.6	74.2	72.3	76.6	93.1	57.3	79.7	43.2	46.2	51.9	31.3	67.8	47.6	64.1
Qwen2.5-VL(7B)	80.0	41.1	70.6	76.6	72.7	48.0	89.2	52.3	64.4	32.4	53.1	55.7	31.3	64.5	49.5	63.8
Phi3.5-vision	35.0	29.4	62.3	43.2	50.9	67.0	40.4	45.2	52.5	37.8	31.7	32.6	24.5	54.1	32.6	45.1
Gemma3(4B)	58.0	52.9	62.3	58.3	60.0	70.0	76.6	33.3	66.1	54.0	29.6	38.5	54.9	54.2	41.1	53.8
Gemma3(27B)	74.0	47.1	67.4	61.3	65.1	66.0	82.9	47.6	69.4	43.2	53.7	55.7	39.2	62.5	51.8	60.7
InternVL2-DA-DriveLM(4B)	50.0	**70.5**	57.2	57.6	57.0	36.0	70.2	52.3	52.5	37.8	43.4	30.7	27.4	58.3	40.5	50.3
OpendriveVLA	74.0	52.9	62.3	75.1	67.9	72.0	89.4	33.3	72.0	43.2	49.7	50.0	35.3	54.1	47.5	60.9
Qwen2.5-VL(7B) (FT, Ours)	**84.0**	64.7	**70.7**	**81.6**	**76.0**	**80.0**	**95.8**	**61.9**	**83.1**	**63.5**	**66.2**	**63.5**	**58.9**	**70.8**	**65.0**	**72.8**

Table 16: Detailed Video QA performance of the MLLMs on WaymoQA. **Bold**: best; underline: second best. The following abbreviations are used: TS (Traffic Signal), SC (Scene Caption), SU (Scene Understanding), PL (Planning), PE (Perception), PP (Perception&Prediction), CF (Counterfactual), SE (Special Event), SR (Spatial Relationship), TG (Temporal Grounding), UN (Uncertainty), FT (Fine-tuning).

scores with stemming enabled. BERTScore is computed with `bert-score` using `roberta-large` with `lang=en` and batch size 32. All metrics are computed at the sample level and averaged within each scene group. We use default package settings unless otherwise specified.

F.4 Prompts

The following system prompt is used for open-ended VLM generation.

Listing 1: System prompt for open-ended VLM generation.

```
You are an autonomous-driving scene question-answering
      assistant.
You will be given one or more images from vehicle cameras.
Answer the question concisely and accurately in 1-2 short
      sentences.
Do NOT list or reference multiple-choice options.
Base your answer only on what is clearly visible in the
      image(s).
If you are uncertain, state the most likely answer with a
      brief reason.
```

The following system prompt is used for GPT-based scoring.

Listing 2: System prompt for GPT-based open-ended answer scoring.

```
You are an expert evaluator specializing in
      autonomous-driving perception and planning.
Your task is to score a model-generated answer against the
      ground-truth answer
for a driving scene question. Score on a 0-100 scale using
      the rubric below.

EVALUATION RUBRIC (total 100 points)

[A] Factual Correctness (0-50 pts)
50 : The predicted action / fact is exactly correct and
      matches the ground truth.
40 : Correct core action, minor wording or detail
      difference (e.g. "slow down" vs "decelerate").
25 : Partially correct -- mentions one correct element
      among multiple, or correct direction but wrong
      maneuver type.
10 : Related to the driving context but the specific
      action/fact is wrong.
 0 : Completely wrong (e.g., "turn right" when GT is "go
      straight").

[B] Completeness (0-30 pts)
30 : All key information from the ground truth is present
      in the prediction.
20 : Most key information present; one minor element is
      missing or imprecise.
10 : Only about half of the key information is captured.
 0 : Critical information (e.g., the main action) is
      missing.

[C] Clarity & Relevance (0-20 pts)
20 : Concise, on-topic, directly answers the question
      without irrelevant content.
12 : Mostly relevant but includes unnecessary filler or
      verbose phrasing.
 6 : Partially on-topic; the answer is confusing or mixes
      correct/incorrect elements.
 0 : Off-topic, incoherent, or refuses to answer.

SCORING NOTES
- Penalise hallucination (adding objects/events not in GT)
      in section [B].
- Reward semantic equivalence even when exact wording
      differs.
- For planning questions, correctness of the final decision
      matters most.
- Do NOT reward verbose answers that pad around a wrong
      core answer.

Respond with ONLY a single integer between 0 and 100. No
      explanation, no prefix.
```

For each example, the user prompt to the GPT judge contains the question, the ground-truth answer, and the predicted answer.

Image QA

Model	Train Dataset	Normal Scene						Others			Safety - Critical							Overall
		TS	SC	SU	PL	PP	Overall	CF	SE	Overall	SU	PE	PR	PL	SC	UN	Overall	
Random Guess	-	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0
Human Experts	-	96.6	97.7	97.7	95.3	94.0	96.1	94.8	96.7	96.3	96.8	98.0	96.4	94.3	98.7	94.9	96.0	96.1
Qwen2.5-VL(7B) (Bai et al., 2025) (Baseline)	-	77.4	77.4	71.2	48.4	63.3	68.1	43.5	58.9	55.6	54.4	55.3	71.0	64.0	43.0	54.6	61.1	62.4
Qwen2.5-VL(7B) (Bai et al., 2025) (FT, Ours)	Image&Video	**84.4**	72.6	72.0	**58.5**	**80.6**	**74.9**	**63.1**	**74.8**	**72.7**	**76.1**	**76.6**	**77.8**	**65.9**	**67.8**	**63.8**	**71.2**	**72.9**
Qwen2.5-VL(7B) (Bai et al., 2025) (FT, Ours)	Image	79.1	78.5	**73.6**	53.9	71.1	71.3	49.1	61.5	59.3	61.6	64.0	73.9	64.9	44.3	54.3	63.6	65.7
Qwen2.5-VL(7B) (Bai et al., 2025) (FT, Ours)	Video	78.6	**80.0**	71.6	50.4	66.7	68.7	46.4	59.1	56.8	59.4	59.3	70.9	62.0	43.7	54.8	61.3	63.4

Video QA

Model	Train Dataset	Normal Scene					Others				Safety - Critical						Overall
		TS	SC	SU	PP	Overall	CF	SE	SR	Overall	SC	SU	PE	TG	UN	Overall	
Random Guess	-	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0
Human Experts	-	95.4	95.7	100.0	98.7	96.9	95.9	100.0	90.5	96.6	97.2	97.2	96.0	94.0	92.5	95.9	96.5
Qwen2.5-VL(7B) (Bai et al., 2025) (Baseline)	-	80.0	41.1	70.6	76.6	72.7	48.0	89.2	52.3	64.4	32.4	53.1	55.7	31.3	64.5	49.5	63.8
Qwen2.5-VL(7B) (Bai et al., 2025) (FT, Ours)	Image&Video	**84.0**	**64.7**	**70.7**	**81.6**	**76.0**	**80.0**	**95.8**	**61.9**	**83.1**	**63.5**	**66.2**	**63.5**	**58.9**	70.8	**65.0**	**72.8**
Qwen2.5-VL(7B) (Bai et al., 2025) (FT, Ours)	Image	78.0	35.3	68.8	76.7	71.4	48.0	91.5	52.4	66.1	33.8	58.0	51.0	36.3	**71.9**	52.9	63.9
Qwen2.5-VL(7B) (Bai et al., 2025) (FT, Ours)	Video	81.0	41.1	69.1	77.9	72.6	51.0	91.5	54.8	67.8	37.9	56.6	59.6	39.2	68.8	54.1	65.1

Table 17: Ablation Study on Image and Video QA of WaymoQA. **Bold**: best; underline: second best. The following abbreviations are used: TS (Traffic Signal), SC (Scene Caption), SU (Scene Understanding), PL (Planning), PE (Perception), PR (Prediction), PP (Perception&Prediction), CF (Counterfactual), SE (Special Event), SR (Spatial Relationship), TG (Temporal Grounding), UN (Uncertainty), FT (Fine-tuning).

Listing 3: User prompt template for GPT-based open-ended answer scoring.

```
[Question]
{question}

[Ground Truth Answer]
{gt_text}

[Predicted Answer]
{pred_text}

Score (0-100):
```

F.5 Results

As shown in Table 18, the WaymoQA fine-tuned Qwen2.5-VL(7B) model achieves the strongest performance across all scene groups on GPTScore and most lexical/semantic metrics. For example, on Overall questions, the fine-tuned model improves GPTScore from 25.5 to 40.4 over the Qwen2.5-VL(7B) base model, and also improves BLEU-4, ROUGE-L, and BERTScore. On Safety-Critical questions, the same model achieves the highest GPTScore and ROUGE-L, indicating that WaymoQA supervision improves free-form safety-critical reasoning rather than only answer-option selection.

At the same time, the relative ordering among pretrained models is less consistent than in the MCQ benchmark. Some models obtain relatively high lexical-overlap scores despite lower GPTScore, while others show similar BERTScore values despite different levels of driving correctness. This supports our evaluation design: human-written MCQs provide the primary objective benchmark, while open-ended evaluation serves as complementary evidence that the observed gains transfer to free-form generation.

G Qualitative Examples of Environmental Augmentation

In the main paper, we briefly describe the use of augmentation to increase robustness to diverse environmental conditions. Here, we provide qualitative examples to illustrate the types of appearance changes used in our augmentation pipeline. Starting from the same base driving scene, we generate variants that simulate different lighting and weather conditions, including dusk, fog, night, overcast daytime, rain, and snow.

These examples are intended to show the visual diversity introduced by augmentation rather than to define a separate benchmark. The goal is to expose the model to appearance shifts that commonly occur in real-world driving, especially those that can affect perception and safety-critical reasoning. By augmenting scenes with variations in illumination, visibility, and weather, we improve robustness to domain shifts while preserving the underlying scene layout and driving semantics.

Figure 6 shows representative augmentation examples generated from a single scene. Although the road geometry, object placement, and traffic context remain unchanged, the visual appearance changes substantially across conditions. This helps the model learn to rely less on superficial appearance cues and better generalize across diverse environments.

Scene Group	Model	GPTScore	B-1	B-2	B-3	B-4	R-1	R-2	R-L	BERTScore
Normal	LLaVA-OneVision(7B)	26.5	8.89	4.74	3.27	2.41	14.3	4.17	12.8	85.3
	InternVL3.5(8B)	30.5	11.2	5.41	3.22	2.05	14.5	4.48	11.9	84.6
	Qwen2.5-VL(7B)	30.5	11.2	5.25	3.23	2.29	14.5	4.48	11.9	84.7
	Qwen2.5-VL(32B)	32.4	10.5	5.12	3.22	2.38	14.0	4.52	11.4	84.1
	Gemma3(27B)	32.1	11.8	5.24	2.89	1.86	15.2	3.85	12.4	84.7
	InternVL2-DA-DriveLM(4B)	28.0	12.2	6.82	4.95	3.81	18.4	6.04	16.5	86.5
	Qwen2.5-VL(7B) (FT)	**45.7**	**28.9**	**18.6**	**13.7**	**11.4**	**39.3**	**18.1**	**36.7**	**89.2**
Others	LLaVA-OneVision(7B)	21.7	3.56	1.78	1.12	0.77	6.71	1.41	6.14	84.3
	InternVL3.5(8B)	26.8	8.99	3.85	2.07	1.31	13.7	3.32	11.0	84.5
	Qwen2.5-VL(7B)	20.7	8.41	3.44	1.82	1.12	12.7	2.74	10.2	84.0
	Qwen2.5-VL(32B)	22.7	8.15	3.09	1.43	0.90	12.7	2.90	9.74	83.7
	Gemma3(27B)	21.6	8.62	3.13	1.53	0.92	12.6	2.34	10.0	84.8
	InternVL2-DA-DriveLM(4B)	26.4	12.2	5.29	2.80	1.82	17.4	4.15	13.7	85.4
	Qwen2.5-VL(7B) (FT)	**44.8**	**20.0**	**13.7**	**7.30**	**4.92**	**29.1**	**15.1**	**27.3**	**87.8**
Safety-Critical	LLaVA-OneVision(7B)	25.5	5.81	2.40	1.23	0.86	11.7	2.81	10.5	85.2
	InternVL3.5(8B)	20.0	9.93	4.25	2.19	1.32	14.3	3.85	11.4	84.8
	Qwen2.5-VL(7B)	23.0	9.96	3.85	1.94	1.12	14.4	3.52	10.2	84.7
	Qwen2.5-VL(32B)	19.9	6.92	2.97	1.54	0.09	11.2	2.69	8.8	82.7
	Gemma3(27B)	21.3	10.8	4.38	2.18	1.32	14.9	3.53	11.7	85.1
	InternVL2-DA-DriveLM(4B)	17.5	11.4	6.66	3.89	2.70	17.5	5.08	15.8	85.8
	Qwen2.5-VL(7B) (FT)	**34.0**	**18.6**	**9.68**	**6.56**	**4.72**	**26.1**	**8.20**	**23.1**	**87.1**
Overall	LLaVA-OneVision(7B)	25.2	6.56	3.17	1.99	1.43	11.7	3.07	10.6	85.1
	InternVL3.5(8B)	25.3	10.2	4.62	2.56	1.59	14.3	3.99	11.5	84.6
	Qwen2.5-VL(7B)	25.5	10.1	4.30	2.41	1.60	14.1	3.74	11.3	84.7
	Qwen2.5-VL(32B)	25.1	8.15	3.09	1.43	0.85	12.7	3.52	10.2	83.7
	Gemma3(27B)	25.5	10.8	4.47	2.33	1.45	14.6	3.53	11.7	85.0
	InternVL2-DA-DriveLM(4B)	23.1	12.7	6.47	4.09	2.96	18.4	6.12	15.7	85.9
	Qwen2.5-VL(7B) (FT)	**40.4**	**22.8**	**13.8**	**9.38**	**7.25**	**31.6**	**13.4**	**29.0**	**88.1**

Table 18: Full open-ended evaluation on WaymoQA. We report GPT-based scoring, BLEU-1/2/3/4, ROUGE-1/2/L, and BERTScore for Normal Scene, Others, Safety-Critical Scene, and Overall questions.



Figure 6: Qualitative examples of environmental augmentation applied to the same driving scene. The scene structure is preserved while lighting, visibility, and weather conditions change across dusk/dawn, foggy, nighttime, overcast, rainy, and snowy variants.

H Failure Case.

Figure 7 illustrates a typical failure pattern of our model. In the first stage of Safety-Critical Reasoning, the model correctly identifies the primary safety-critical object (a person and a dog next to the parked vehicle), predicts that the person will remain standing near the car, and concludes that the ego vehicle should wait until the person either walks away or re-enters the vehicle. Thus, basic perception, prediction, and planning around the first hazard are handled reliably.

However, the second stage exposes a systematic weakness when directions are defined in a *relative* coordinate frame. The additional safety-critical object is an oncoming blue vehicle, described in the ground truth as "driving in the opposite lane to

the left of the white vehicle in front" and "attempting to drive straight ahead to the left of the white vehicle in front." Our model instead predicts that the blue vehicle is on the right side of the white vehicle and moving to the right. In other words, it fails to reinterpret "left" and "right" from the other vehicle's perspective and mistakenly projects them into the ego-centric frame. This misinterpretation then propagates to downstream sub-questions: the model's subsequent plan for "addressing the first risk while also addressing the second risk" is inconsistent with the true geometry of the scene, leading to an unsafe recommendation. A similar confusion appears in the normal reasoning question at the bottom of the figure, where the model again flips the turning direction of the rear vehicle when the description is given from that vehicle's viewpoint.

I Extending WaymoQA

Our current benchmark evaluates *perception*, *prediction*, and *planning* with separate questions. For example, in the scenario shown in Fig. 7, we ask (i) which object is safety-critical, (ii) how that object will move, and (iii) how the ego vehicle should drive safely. While this decomposition reveals which sub-skill fails, it does not explicitly enforce that the model maintains a coherent reasoning chain from perception to planning. A model can sometimes answer each sub-question in isolation by pattern matching, without truly understanding how its perception of agents constrains future motion and, in turn, ego decisions.

Because WaymoQA contains rich, temporally aligned annotations, we can instead reorganize these items into *compound* questions that evaluate the full decision process. For instance, rather than three separate questions, a single multi-part query could ask: (1) which agents are safety-critical in the current scene, (2) how each of them is likely to move over the next few seconds, and (3) which maneuver the ego vehicle should take to reduce overall risk. The ground truth already contains all of this information, so the model's answer can be graded both holistically and at the level of partial credit: a model that correctly identifies the primary hazard but mispredicts the second agent's motion would receive credit for perception but be penalized on prediction and planning. Such structured supervision directly ties mistakes in downstream planning to specific upstream errors, e.g., misinterpreting a relative left/right description for an oncoming vehicle.

We view these extensions as a natural next step for WaymoQA. By transforming our existing annotations into multi-stage, chain-structured questions with explicit partial scoring, future work can train and test models on *complete* safety-critical reasoning pipelines rather than isolated sub-tasks, and thus more directly target the types of compounding failures observed in our error analysis.



1st Stage Safety-Critical Reasoning

Question: What is a safety-critical object?

GT: B. A person and a dog standing on the left side of a vehicle parked on the right shoulder ahead.

Ours: B. A person and a dog standing on the left side of a vehicle parked on the right shoulder ahead.

Question: How do you think a safety-critical object will move?

GT: D. It seems like it will stand (bending over and standing while fiddling with something inside the car)

Ours: D. It seems like it will stand (bending over and standing while fiddling with something inside the car)

Question: What should Ego do to drive safely?

GT: B. Ego must wait until this person moves on foot or gets back into the vehicle

Our: B. Ego must wait until this person moves on foot or gets back into the vehicle..

2nd Stage Safety-Critical Reasoning

Question: Is there any other safety-critical object besides the first safety-critical object? If so, how will it be moved?

GT: B. A blue vehicle is driving in the opposite lane to the left of the front. It is attempting to drive straight ahead to the **left** of the white vehicle in front.

Ours: C. A blue vehicle is driving in the opposite lane to the right of the front. It is attempting to drive straight ahead to the **right** of the white vehicle in front.

Question: How should I drive to address the first risk while also addressing the second risk?

GT: A. Ego must wait until this person moves on foot or gets back into the vehicle.

Ours: D. It is safe to stop, but it is also possible to move to the left.

Normal Reasoning

Question: Based on the current appearance of vehicle, how do they intend to move?

GT: C. The black car in the rear is turning **left** at the intersection.

Ours: A. The black car in the rear is turning **right** at the intersection.

Figure 7: **Failure case on two–stage Safety-Critical Reasoning in WaymoQA.**