Measuring Massive Multimodal Understanding and Reasoning in Open Space

https://open-space-reasoning.github.io

Shangding Gu[1], Xiaohan Wang[2], Donghao Ying[1], Haoyu Zhao[3], Runing Yang[4], Boyi Li[1,5], Ming Jin[4], Marco Pavone[2,5], Serena Yeung-Levy[2], Jun Wang[3], Dawn Song[1†], Costas Spanos[1†]

[1]UC Berkeley [2]Stanford [3]UCL [4]Virginia Tech [5]Nvidia [†]Equally advise

Abstract

The increasing sophistication of multimodal models necessitates benchmarks that can rigorously evaluate their understanding and reasoning in complex, safety-pertinent, open-world scenarios. This study introduces M4R (Measuring Massive Multimodal Understanding and Reasoning), a large-scale benchmark uniquely designed to assess reasoning capabilities across diverse open spaces, comprehensively covering land, air, and water environments. M4R comprises approximately 2,000 videos and over 19,000 human-annotated question-answer pairs. These videos, varying in length (short, medium, long) and presenting tasks of tiered difficulty (interval-based choices and accuracy-based choices), encompass distinct operational domains: the land-based scenarios primarily focus on traffic environments, particularly traffic collisions and accident cases; the air-based scenarios center on airplane navigation; and the water-based scenarios involve ship movements. M4R systematically evaluates models on temporal reasoning, spatial understanding, and intent inference within these dynamic contexts. By providing a unified platform across this broad spectrum of domains, M4R aims to drive the development of safer, more robust, and generalizable AI systems. Benchmarking state-of-the-art multimodal models on our dataset reveals that even leading models, such as ChatGPT-4o and Gemini-2.5 Pro, achieve only around 30% average accuracy on the hard-level tasks, highlighting the significant challenges that remain in open-space multimodal reasoning. The code, leaderboard, and dataset are available at: `https://open-space-reasoning.github.io`.

1 Introduction

As artificial intelligence (AI) continues to evolve, large multimodal models have shown impressive capabilities across vision, language, and video domains. However, significant challenges remain in deploying these models for real-world, safety-critical applications such as autonomous driving, robotics, and aerial or maritime operations. While multimodal models demonstrate remarkable performance in constrained or simulated environments, their robustness and depth of understanding in high-stakes, dynamic scenarios are still far from sufficient.

In particular, deployment in mission-critical domains requires rigorous evaluation of models' understanding and reasoning abilities under real-world conditions that involve uncertainty, physical interactions, and causal dependencies. While recent benchmarks have advanced evaluation in specific

Technical Report.

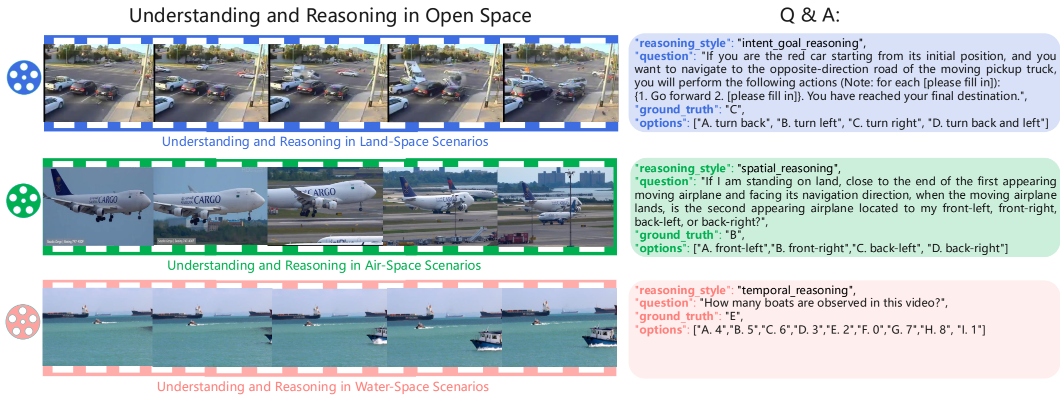


Figure 1: Examples of multimodal Understanding and Reasoning in Open-Space Scenarios

facets like temporal understanding (e.g., MVBench [23], REXTIME [6]) or domain-specific knowledge (e.g., MMMU [46], DriveLM [33]), there remains a paucity of unified platforms that assess reasoning across the combined spectrum of land, air, and water operations. To address this, our work defines *open space* as unstructured or semi-structured outdoor environments characterized by high variability, dynamic interactions, and minimal physical boundaries. This includes **air space** (e.g., airplane navigation), **water space** (e.g., ship and boat movements), and **land space** (e.g., road traffic involving diverse vehicle types). These settings inherently involve complex temporal dependencies, causal relationships, and real-world physical constraints, demanding advanced, robust reasoning capabilities for genuine open-world understanding.

We introduce M4R, a comprehensive evaluation framework. Specifically, M4R focuses on reasoning across the aforementioned land traffic, airspace, and waterway domains—settings where safety, perception, and decision-making are deeply interdependent. Unlike benchmarks focusing on isolated skills or single domains, M4R challenges models on several key reasoning capabilities: *temporal-causal reasoning* (understanding event sequences and causality over extended periods); *spatial understanding* (comprehending dynamic spatial relationships and multi-agent trajectories); *intent and goal planning/inference* (deducing agent intentions and goals), which includes *complex strategic & counterfactual reasoning* (assessing understanding of higher-order strategies, action implications, and 'what-if' scenarios). Several representative examples from M4R are illustrated in Figure 1. By systematically probing these capabilities across diverse safety-pertinent scenarios, M4R provides a framework for assessing progress towards AI systems that can reliably operate in the real world.

Our key contributions are summarized as follows:

- **Unified Open-World Evaluation Suite:** We introduce M4R, a large-scale, video-based benchmark uniquely covering land traffic, airspace, and waterway scenarios to provide a comprehensive assessment of multimodal reasoning across these distinct yet complementary safety-critical open spaces.
- **Reasoning-Centric Evaluation:** M4R systematically evaluates critical reasoning facets including temporal-causal understanding, dynamic spatial awareness, intent and goal reasoning, within dynamic and physically grounded settings.
- **Real-World Limitations and Safety Gaps:** We highlight limitations in current AI systems' reasoning performance in open-space domains (e.g., autonomous driving, aviation, and maritime environments), and provide a challenging testbed to drive the development of safer and more robust multimodal AI systems.

2 Related Work

2.1 General Multimodal Understanding Benchmarks

Recent years have witnessed growing interest in video understanding benchmarks. Foundational video question-answering (QA) efforts include MSR-VTT [44] and Next-QA [42]. More recently, MVBench [23], with its 20 diverse temporal tasks derived from static images, and MLVU [49] have expanded video QA capabilities across multiple domains. The challenge of long-form video understanding has seen contributions from benchmarks such as EgoSchema [27], Video-LLaVA [9], MovieChat [35], and LongVideoBench [41]. Parallelly, video captioning benchmarks such as Aurora-Cap [5], HiCM2 [20], and LongCaptioning [40] focus on generating detailed textual descriptions.

A significant trend is the push for more rigorous temporal and causal reasoning. REXTIME [6], for instance, probes the linking of causally related events across separate video segments. For multi-domain understanding, MMWorld [16] evaluates models across diverse disciplines, requiring explanations and counterfactuals. Furthermore, LVBench [39] integrates video inputs for QA. Beyond video, reasoning from static images is explored by MME [19] (including CoT extensions), MMMU [46] (evaluating expert-level multi-discipline reasoning), and benchmarks for mathematical reasoning like Dynamath [51] and MultiModal-MATH [50]. For academic content, Video-MMLU [37] offers a large-scale lecture video benchmark.

While these diverse benchmarks significantly advance specific aspects of multimodal understanding—be it general video comprehension, temporal analysis, long-form narrative understanding, captioning, or static image reasoning—they often do not provide a framework for unified evaluation across land, air, and maritime open-space environments, nor the specific blend of complex reasoning (including strategic and intent-based inference) that M4R is designed to evaluate within these contexts.

2.2 Safety-Critical Multimodal Understanding Benchmarks

Evaluating models in safety-critical domains, where reasoning under uncertainty is vital, is an emerging focus. Initial efforts addressed static image safety [24], model robustness against adversarial attacks (e.g., FigStep [11], JailBreakV [26]) [32, 29], or indoor robotics [45].

Autonomous driving has been a major driver of safety-critical research. Foundational datasets such as nuScenes[1] and Waymo Open Dataset[2], along with language-integrated efforts such as DriveLM and DriveVLM [33, 38], are closely related to M4R's goals due to their real-world video and safety considerations. However, a key motivation for M4R was that these traditionally emphasized perception and planning, with less focus on deep safety-critical reasoning for tasks such as accident cause analysis or complex decision-making. Other specialized benchmarks tackle related issues such as video anomaly detection (e.g., VANE-Bench [10]).

While advancements continue in specialized video reasoning and domain-specific safety evaluations, existing benchmarks still largely focus on single operational domains. Critically, they often lack sufficient coverage of high-risk scenarios such as traffic collisions, ship navigation, and airplane takeoff/landing events across combined land, air, and water settings. A unified platform to consistently evaluate robust, generalizable reasoning (e.g., temporal-causal, spatial, intent, and strategic analysis) across these diverse, safety-critical open spaces also remains absent. To address this specific void, M4R distinctively incorporates these challenging high-risk scenarios from all three domains. The reliability of its complex reasoning evaluation is ensured as all annotations were generated by highly educated annotators (at least Master's degree). M4R thus provides a much-needed testbed for fostering robust, adaptable AI capable of open-world understanding.

3 Understanding and Reasoning in Open Space

3.1 Open Space Settings

We design the benchmark around three types of open-space environments: **land space**, focusing primarily on traffic accident understanding and reasoning; **air space**, centered on airplane takeoff and landing scenarios; and **water space**, which emphasizes ship navigation understanding and reasoning. Within each environment, we construct tasks that evaluate models across three key reasoning dimensions: dynamic temporal reasoning, spatial reasoning, and intent and goal reasoning. Representative examples for each reasoning type are illustrated in Figure 2.

For each reasoning style, we design tasks with varying levels of difficulty using two formats: *interval-based choices* and *accuracy-based choices*. Easy tasks provide approximately 3 coarse-grained

[1] `https://www.nuscenes.org/`
[2] `https://waymo.com/open/`

Figure 2: Examples of reasoning question settings in M4R across three key reasoning types: *Temporal Reasoning*, which involves understanding event sequences and motion over time; *Spatial Reasoning*, which focuses on relative positioning and orientation in space; and *Intent Reasoning*, which evaluates understanding of goal-directed behaviors and decision-making in dynamic environments.

interval choices, medium tasks offer 6 intermediate-level intervals, and hard tasks present 12 fine-grained discrete options that require an exact match with the correct answer. The number of tasks across the three difficulty levels is evenly distributed, with each comprising one-third of the total. In all cases, the model must select a single best answer, enabling the benchmark to assess performance under increasing levels of precision and ambiguity.

Land Space In our land-space setting, we include a comprehensive range of traffic scenarios, encompassing diverse collision events under varying weather conditions such as snow, rain, and sunshine, as detailed in Table 1. Specific examples of these scenarios are illustrated in Figure 3, and more detailed examples are provided in Appendix B. To enhance contextual diversity, we incorporate multiple camera perspectives—including ego-centric and third-person views—particularly for accident scenes. The dataset features incidents involving a wide variety of vehicle types, including buses, motorcycles, sedans, and several categories of trucks, across different road environments such as highways, freeways, and rural roads. The associated questions are designed to evaluate models across multiple reasoning dimensions, including temporal-causal understanding, spatial reasoning, and intent and goal planning. The original land-space video datasets are sourced from [4, 31], which primarily collected videos from YouTube and other public internet platforms.

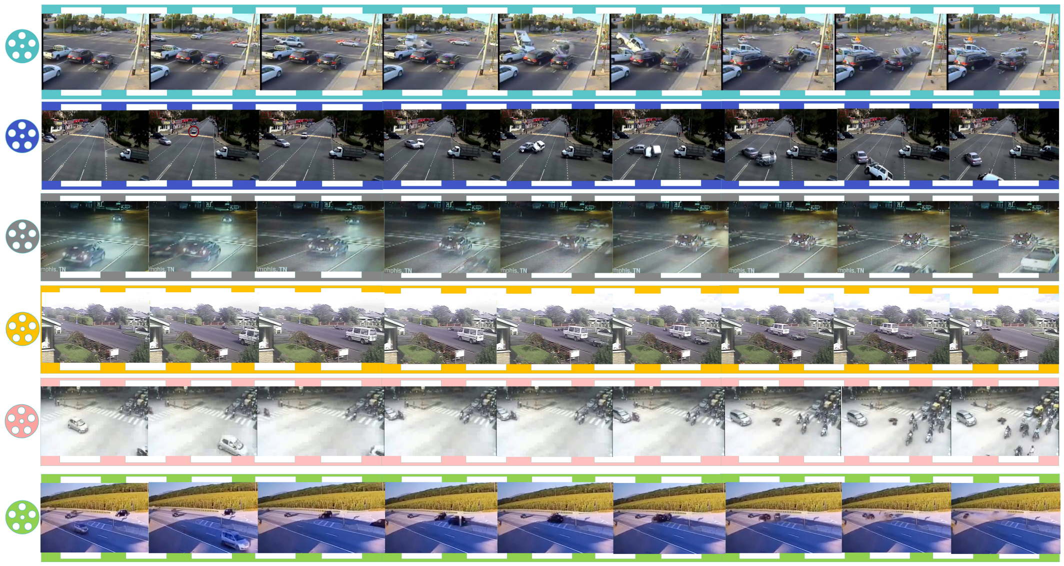


Figure 3: Land-space traffic accident scenarios for open-space video understanding and reasoning include intersection collisions, urban road accidents, nighttime incidents, rural road accidents, snow-covered road collisions, and freeway accidents.

Table 1: Overview of traffic accident scenarios in our benchmark, covering diverse road environments, weather conditions, and involved traffic participants.

Index	Categories
Road Environments:	Intersection, Highway, Freeway, Rural Road, Tunnel, Urban Road, Bridge, Parking Lot
Weather Conditions:	Snow, Rain, Sunshine, Cloudy, Foggy, Windy
Involved Participants:	Sedan, SUV, Bus, Truck, Motorcycle, Bicycle, Van, Pickup, Trailer, Pedestrian

Air Space In airspace scenarios, we primarily focus on *takeoff* and *landing* events, emphasizing the analysis of airplane navigation directions and perceptual understanding. Airplanes represent a largely unexplored domain in large multimodal research, despite their significant real-world impact. Our benchmark investigates various aspects of airplane behavior, including differences in navigation patterns, aircraft sizes, and motion dynamics across different types of airplanes. These scenarios also incorporate videos of varying lengths and are designed to evaluate models on multiple reasoning dimensions, including spatial reasoning, temporal reasoning, and intent and goal inference. We further assess model performance across different difficulty levels using both interval-based and accuracy-based multiple-choice formats. The airspace videos are sourced from publicly available footage, including references such as [3], [4], and [5].

Water Space We include videos from both **river** and **ocean** scenarios, featuring varying video lengths and difficulty levels. The dataset encompasses a diverse range of watercraft, including different types of boats and ships, under a broad set of navigation conditions. Despite their real-world importance, river and ocean environments remain underexplored in the context of large multimodal models. To address this gap, we evaluate model performance across multiple reasoning styles—temporal, spatial, and intent and goal reasoning—using video-based tasks of varying durations and difficulty levels. Task difficulty is controlled through both interval-based and accuracy-based multiple-choice formats. The water-space videos are sourced from publicly available datasets, including [14, 28].

3.2 Dataset Analysis

This benchmark includes approximately 2,000 videos and 19,000 human-annotated question-answer pairs, covering a wide range of reasoning tasks. All annotations were performed by highly educated

[3] `https://www.youtube.com/watch?v=i6CrbqeksJ8`

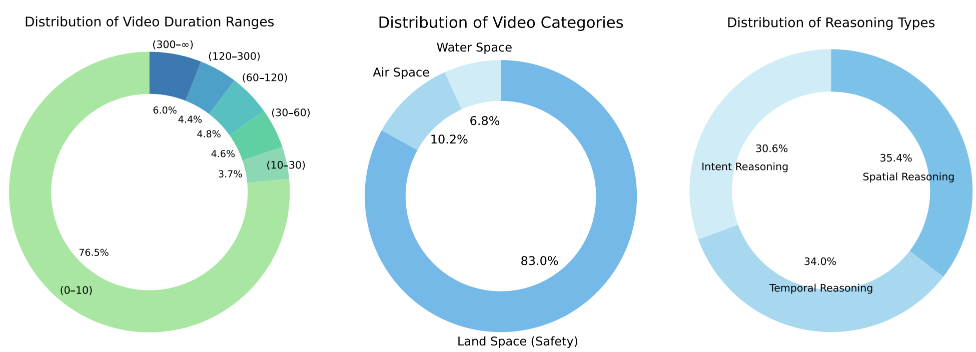
[4] `https://www.youtube.com/watch?v=k5yvzTw08K8`

[5] `https://www.youtube.com/watch?v=Bt9tpiAmTs8`

annotators, each holding at least a master's degree in engineering-related fields such as mathematics or computer science. The dataset features a variety of video lengths, categories, and frame counts, and spans three primary open-space reasoning scenarios: **land space**, **water space**, and **air space**. An overview of the dataset's characteristics is shown in Figure 4, which illustrates the distributions of video duration, domain coverage, and reasoning styles. During annotation, we first design the hard-level tasks and label each question with the ground-truth answer. Based on these, we then construct the medium and easy tasks. The primary differences between difficulty levels lie in the number and types of answer choices. Details of the annotation procedure and difficulty levels are provided in Appendix B.

Specifically, **(a) Video Length:** A substantial portion of the videos (76.5%) are short, with durations under 10 seconds. The remaining videos are distributed across longer intervals: 10–30 seconds (3.7%), 30–60 seconds (4.6%), 60–120 seconds (4.8%), 120–300 seconds (4.4%), and over 300 seconds (6.0%). This distribution reflects a strong emphasis on short, dynamic scenarios that test rapid perception and reasoning. **(b) Video Categories:** The benchmark spans three open-space domains. Land space, which primarily involves traffic and safety-related scenarios, comprises 83.0% of the videos. Air space accounts for 10.2%, and water space makes up 6.8%. This distribution highlights both the practical importance of land-based reasoning and the inclusion of underrepresented domains such as maritime and aviation environments. **(c) Reasoning Styles:** M4R supports three major reasoning types, with a relatively balanced distribution: *spatial reasoning* (35.4%), *temporal reasoning* (34.0%), and *intent reasoning* (30.6%). This design ensures comprehensive evaluation across key dimensions essential for real-world multimodal understanding.

Overall, the dataset provides a rich and diverse collection of real-world video scenarios across multiple modalities and time scales, offering a robust benchmark for evaluating multimodal understanding and reasoning in open-space environments.



(a) Video duration. (b) Video categories. (c) Reasoning types.

Figure 4: Distribution of video and task properties in the M4R benchmark.

3.3 Comparison with Existing Benchmarks

Table 2 provides a comparative analysis of M4R alongside existing evaluation benchmarks for multimodal models. Most benchmarks primarily focus on assessing the multimodal reasoning capabilities of multimodal models [15, 36, 49]; however, a significant limitation is the prevalent oversight of safety considerations. While a few recent benchmarks have begun to evaluate safety aspects of multimodal models [50, 24], they often do not incorporate video question-answering data. However, single-frame capture, in most cases, can introduce uncertainties in reasoning and is insufficient for adequately assessing multimodal models' capabilities in handling safety issues. In contrast, our M4R introduces a large-scale curated collection of video question-answer pairs that specifically focus on open-space traffic reasoning in real-world safety-related scenarios. Comprising 2,000 carefully selected videos and 19,000 reasoning question-answer pairs, the M4R features a size competitive with existing benchmarks, thus highlighting the comprehensiveness of our evaluation set.

Table 2: **Benchmark comparison** for multimodal understanding and reasoning tasks.

Dataset	Safety	Traffic	Annotation	Real-World	Scenarios	# Video	Ave. Duration (s)	Question-answering Number	Type
MovieChat-1K [36]	×	×	Human	✓	General	1,000	564	13,000	Open-ended
MMWorld [15]	×	×	Human	✓	General	1,910	107	6,627	Multiple-choice
MLVU [49]	×	×	Human	✓	General	1,730	930	3,102	Multiple-choice
MVBench [1]	×	×	Human & LLM	✓	General	4,000	16	4,000	Multiple-choice
LongVideoBench [41]	×	×	Human	✓	General	3,763	473	6,678	Multiple-choice
TempCompass [25]	×	×	Human & LLM	✓	General	410	< 30	7,540	Multiple-choice
VSI-Bench [45]	×	×	Human	✓	Embodied	288	50-100	5,000	Multiple-choice
Video-MMMU [17]	×	×	Human & LLM	×	Professional	300	506	900	Multiple-choice
Video-MMLU [37]	×	×	Human & LLM	×	Professional	1,065	109	15,746	Open-ended
DriveBench [43]	✓	✓	Human & LLM	✓	Autonomous Driving	×	×	19,200	Multiple-choice
DriveLM [34]	✓	✓	Human	✓	Autonomous Driving	×	×	15,480	Open-ended
nuScenes-QA [30]	×	✓	Human	✓	Autonomous Driving	×	×	83,337	Open-ended
MSSBench [50]	✓	×	Human & LLM	✓	General	×	×	1960	Open-ended
MMSBench [24]	✓	×	LLM	✓	General	×	×	5040	Open-ended
M4R (ours)	✓	✓	Human	✓	General	2000	56	19,000	Multiple-choice

Table 3: Multimodal understanding and reasoning evaluation using M4R in open space, across Land, Water, and Air domains (columns = difficulty levels).

Models	Size	Hard				Medium				Easy			
		Avg.	Temp.	Spatial	Intent	Avg.	Temp.	Spatial	Intent	Avg.	Temp.	Spatial	Intent
GPT 4o [18]	-	22.21	24.92	27.14	13.80	41.21	44.89	47.03	28.19	45.01	55.33	38.08	43.72
Gemini 2.5 pro [13]	-	**31.01**	38.18	30.08	25.20	41.07	41.31	48.33	33.06	**59.36**	61.16	54.51	58.09
Gemini 2.5 flash think [12]	-	28.52	31.74	30.33	26.36	**41.45**	46.83	45.89	35.51	53.14	58.70	55.86	48.47
Gemini 2.5 flash no-think [12]	-	24.33	24.80	30.41	20.53	40.36	41.97	42.93	33.61	50.52	52.16	51.30	44.41
Gemini 1.5 pro [8]	-	19.07	22.53	21.57	17.25	37.13	40.69	43.81	31.06	48.05	53.22	47.85	45.37
Claude 3.5 [2]	-	28.89	32.84	29.18	23.41	37.99	36.46	47.34	31.09	50.14	53.28	48.51	46.40
InternVL2.5 [7]	26B	22.45	25.33	27.42	12.64	36.39	37.85	47.51	27.55	55.08	58.41	53.46	44.45
InternVL2.5 [7]	8B	20.39	21.30	29.41	11.42	35.44	39.85	51.07	18.98	51.03	53.64	54.52	42.20
InternVL2.5 [7]	4B	17.31	17.39	23.04	13.13	36.53	31.21	45.36	32.68	48.93	46.55	52.31	43.65
LLaVA Next [21]	32B	17.83	11.28	26.09	10.10	21.07	13.57	33.08	14.24	35.32	31.22	40.09	34.34
LLaVA Video [48]	7B	17.35	13.02	27.49	10.18	24.04	19.33	30.50	19.72	30.44	29.41	34.12	31.64
LLaVA OneVision [22]	7B	14.27	9.55	24.74	10.15	17.76	17.81	24.71	17.12	31.10	29.46	33.78	29.88
Qwen2.5 VL [3]	32B	19.39	13.19	27.85	14.05	29.93	23.34	41.94	25.82	48.35	50.68	47.82	44.97
Qwen2.5 VL [3]	7B	20.34	12.31	28.40	15.48	28.79	22.18	34.64	22.89	37.97	38.87	33.20	36.45

4 Experiments

In our experiments, we build upon the `lmms-eval` framework [47] as the foundation for our benchmark and extend it to support the specific requirements of M4R. We conduct comprehensive evaluations to assess the performance of SOTA multimodal models across diverse open-space scenarios.

4.1 Comparative Evaluation of Multi-Modal Models in M4R

Table 3 presents the performance of multi-modal models across the Land, Water, and Air domains in the OpenRBench benchmark. Results indicate that SOTA proprietary models, such as Gemini 2.5 Pro, Genimi 2.5 flash, and GPT-4o, outperform open-source counterparts, particularly on all tasks. Gemini 2.5 Pro achieves the highest average score (59.36) on easy tasks and maintains strong performance in hard and medium settings. Genimi 2.5 flash, while slightly behind on easy tasks, leads in medium difficulty with a score of 41.45. However, all models, including the strongest ones, exhibit notable performance degradation on hard tasks, with average scores falling below 35.

Among open-source models, all achieve only around 20 accuracy on hard tasks. The Qwen-based models generally outperform LLaVA-based models, with one possible reason being architectural differences. For instance, Qwen2.5-VL employs a dynamic vision encoder capable of handling variable spatial-temporal resolutions, whereas LLaVA-Video uses a fixed-resolution vision encoder. These design choices likely contribute to their differing performance across reasoning tasks. However, since each model is trained on distinct proprietary datasets, it remains challenging to disentangle and isolate the impact of architecture alone. These results highlight persistent limitations in complex real-world reasoning, including temporal, spatial, and intent understanding, and highlight the need for more robust and generalizable multi-modal systems for open-space environments.

4.2 Performance Analysis for Each Space Task

Land Space Evaluation: Table 4 presents the evaluation results of multi-modal models on the Land domain of OpenRBench, categorized by task difficulty (Easy, Medium, Hard) and reasoning type (Temporal, Spatial, Intent). Among the models evaluated, Gemini 2.5 Pro consistently performs best in the Easy setting with an average score of 57.90, and remains competitive in Medium (40.57) and Hard (31.06) tasks. GPT-4o shows strong performance in Medium (43.05) and Easy (44.17) tasks, leading in temporal and spatial reasoning, but struggles more with hard tasks (25.82). Across all models, performance declines significantly with increasing task difficulty, particularly in intent reasoning under the Hard setting—highlighting a persistent challenge for current models. Overall, proprietary models (e.g., Gemini, GPT-4o) outperform open-source counterparts, though no model achieves robust performance across all difficulty levels and reasoning types.

Table 4: Understanding and reasoning evaluation for M4R in Land Space domain.

Models	Size	Hard				Medium				Easy			
		Avg.	Temp.	Spatial	Intent	Avg.	Temp.	Spatial	Intent	Avg.	Temp.	Spatial	Intent
GPT 4o [18]	-	25.82	29.61	31.38	13.21	**43.05**	47.63	48.59	31.83	44.17	54.45	36.01	43.67
Gemini 2.5 pro [13]	-	**31.06**	38.75	37.54	23.46	40.57	39.13	47.22	30.33	**57.90**	58.24	56.23	55.52
Gemini 2.5 flash think [12]	-	29.90	34.52	36.57	23.00	39.50	45.18	45.61	29.76	48.93	58.35	51.21	37.78
Gemini 2.5 flash no-think [12]	-	23.80	24.43	33.04	17.55	36.67	41.21	43.86	25.44	46.89	50.92	50.24	35.19
Gemini 1.5 pro [8]	-	17.79	20.90	20.72	15.81	35.98	39.05	41.11	28.75	47.00	56.01	45.68	40.25
Claude 3.5 [2]	-	30.82	35.04	31.65	22.91	37.93	36.39	46.36	32.63	51.08	53.32	47.01	48.93
InternVL2.5 [7]	26B	23.92	31.00	29.75	11.50	35.42	41.75	43.00	22.75	56.33	61.00	56.25	46.50
InternVL2.5 [7]	8B	21.25	24.50	31.25	10.50	34.83	42.25	48.25	14.50	52.34	55.50	57.00	42.50
InternVL2.5 [7]	4B	17.50	19.50	25.50	12.00	35.33	34.00	41.25	26.50	48.00	46.00	51.50	43.50
LLaVA Next [21]	32B	19.34	13.50	24.50	11.00	21.83	15.50	31.25	14.00	37.09	27.25	41.75	35.00
LLaVA Video [48]	7B	19.67	15.00	31.25	12.00	25.42	22.00	32.50	22.50	30.58	31.00	32.25	34.00
LLaVA OneVision [22]	7B	13.83	8.50	21.75	12.00	16.67	20.50	19.00	17.00	30.83	29.75	32.25	29.00
Qwen2.5 VL [3]	32B	23.33	18.00	29.50	18.00	27.99	25.75	38.50	23.50	45.67	53.00	45.50	41.25
Qwen2.5 VL [3]	7B	23.42	18.25	30.50	20.75	32.17	30.50	37.00	24.00	43.58	44.00	38.50	42.75

Air Space Evaluation: Table 5 reports the evaluation results for multi-modal models in the Air Space domain of OpenRBench. The results are broken down by task difficulty (Easy, Medium, Hard) and reasoning types (Temporal, Spatial, Intent). Gemini 2.5 Pro stands out with the strongest overall performance, achieving the highest average scores across all difficulty levels, including 31.86 (Hard), 41.21 (Medium), and 55.74 (Easy). It particularly excels in intent reasoning, reaching up to 61.17 in the Easy setting. GPT-4o also performs competitively, especially on Easy tasks (40.72) and intent reasoning (39.67), though it lags behind Gemini on harder examples. Open-source models such as InternVL2.5 and Qwen2.5 show moderate success in temporal reasoning but consistently underperform in intent reasoning. Overall, the trend mirrors that of the Land domain: performance declines significantly as difficulty increases, with the largest drop occurring in temporal and intent reasoning tasks. These results emphasize the challenges multi-modal models face in reliably operating in dynamic, real-world Air Space scenarios.

Table 5: Understanding and reasoning evaluation for M4R in Air Space domain.

Models	Size	Hard				Medium				Easy			
		Avg.	Temp.	Spatial	Intent	Avg.	Temp.	Spatial	Intent	Avg.	Temp.	Spatial	Intent
GPT 4o [18]	-	18.02	12.21	29.77	15.46	30.53	31.33	40.83	31.83	40.72	37.83	37.00	39.67
Gemini 2.5 pro [13]	-	**31.86**	34.26	21.56	34.25	41.21	44.08	38.25	53.50	**55.74**	59.72	47.17	61.17
Gemini 2.5 flash think [12]	-	25.78	26.00	18.00	34.00	39.78	39.33	32.00	48.00	50.67	49.33	40.00	62.00
Gemini 2.5 flash no-think [12]	-	25.44	25.33	22.00	28.00	**49.67**	43.33	30.00	52.00	50.78	49.33	36.00	60.00
Gemini 1.5 pro [8]	-	22.88	19.15	24.75	22.25	36.21	32.83	49.50	32.00	43.89	40.56	41.89	49.67
Claude 3.5 [2]	-	24.31	16.55	32.30	23.00	36.44	32.60	47.79	33.33	41.03	37.56	41.61	45.33
InternVL2.5 [7]	26B	18.60	17.75	26.50	12.00	20.14	26.31	23.31	46.83	32.11	36.31	34.25	46.42
InternVL2.5 [7]	8B	18.71	14.80	29.75	10.00	23.73	30.42	32.92	46.00	37.86	40.33	36.50	40.00
InternVL2.5 [7]	4B	15.14	14.25	16.75	13.13	24.41	27.00	28.25	46.75	38.31	39.64	38.39	41.25
LLaVA Next [21]	32B	18.23	8.98	35.08	10.15	20.71	17.47	37.33	21.67	28.60	32.69	34.36	34.67
LLaVA Video [48]	7B	15.56	8.48	25.80	9.00	20.35	20.25	25.83	21.33	29.62	30.94	30.97	30.00
LLaVA OneVision [22]	7B	15.76	11.00	26.75	9.50	19.81	19.84	23.83	20.83	29.62	30.94	30.97	30.00
Qwen2.5 VL [3]	32B	16.35	3.43	31.08	13.75	35.85	29.00	27.17	43.67	51.73	52.33	40.61	54.44
Qwen2.5 VL [3]	7B	16.38	1.16	30.00	16.00	28.70	22.61	30.33	25.83	38.92	35.78	36.39	30.00

Water Space Evaluation: Table 6 presents evaluation results for multi-modal models in the Water Space domain of OpenRBench. The evaluation is broken down by task difficulty (Easy, Medium, Hard) and reasoning type (Temporal, Spatial, Intent). Gemini 2.5 Pro again leads overall performance, achieving the highest average scores in both Easy (60.92) and Hard (28.11) tasks, and competitive results in Medium (40.92). It demonstrates particularly strong spatial and intent reasoning capabilities. GPT-4o performs well in Medium (37.30) and Easy (47.16) settings, but struggles more with hard tasks (19.97). Open-source models such as InternVL2.5 and Qwen2.5 show varying levels of competence, particularly in temporal reasoning. As with other domains, all models experience a marked performance drop on hard tasks, especially in intent reasoning. These results reflect the continued difficulty of multi-modal reasoning in dynamic, ambiguous environments like rivers and oceans, reinforcing the need for more advanced AI systems.

Table 6: Understanding and reasoning evaluation for M4R in Water Space domain.

Models	Size	**Hard**				**Medium**				**Easy**			
		Avg.	Temp.	Spatial	Intent	Avg.	Temp.	Spatial	Intent	Avg.	Temp.	Spatial	Intent
GPT4o [18]	-	19.97	22.62	21.29	14.73	37.30	39.48	50.53	20.84	47.16	63.06	38.68	41.69
Gemini 2.5 pro [13]	-	**28.11**	33.38	22.06	26.80	40.92	44.30	56.28	29.45	60.92	68.47	52.68	58.04
Gemini 2.5 flash think [12]	-	27.17	31.39	29.15	23.42	**46.72**	52.02	56.24	35.00	**62.01**	65.20	74.48	52.74
Gemini 2.5 flash no-think [12]	-	24.76	25.61	38.20	18.02	42.27	42.37	50.28	32.30	59.15	57.84	68.01	51.06
Gemini 1.5 pro [8]	-	25.48	31.25	23.57	21.97	41.86	48.64	50.01	41.17	49.84	47.47	50.30	50.02
Claude 3.5 [2]	-	24.14	23.67	20.77	26.06	39.26	40.07	53.80	26.67	50.27	58.37	52.46	39.70
InternVL2.5 [7]	26B	22.35	17.68	25.19	22.01	41.01	25.68	60.34	34.78	52.42	55.55	51.60	43.28
InternVL2.5 [7]	8B	21.98	13.74	27.65	21.21	41.01	33.81	60.90	25.26	51.51	57.54	51.19	46.09
InternVL2.5 [7]	4B	20.92	17.01	24.68	21.60	44.13	27.18	62.23	44.04	53.28	52.10	55.76	44.42
LLaVA Next [21]	32B	13.85	7.96	27.13	7.98	20.18	10.84	33.10	16.68	35.00	34.48	39.46	33.38
LLaVA Video [48]	7B	13.45	9.70	21.59	7.10	22.14	19.81	29.13	18.95	30.31	23.56	37.22	30.00
LLaVA OneVision [22]	7B	15.00	9.42	27.25	8.42	22.59	16.27	32.09	18.29	32.95	29.67	37.08	31.44
Qwen2.5 VL [3]	32B	12.99	7.97	23.63	7.37	33.25	19.69	50.00	29.72	52.04	45.12	56.49	43.05
Qwen2.5 VL [3]	7B	13.76	7.02	26.33	8.00	26.10	18.94	28.36	24.67	30.17	34.70	20.74	34.95

4.3 Performance Analysis for Each Space Task with Video Length and Scenarios

Table 7: Evaluation of M4R in the **Land Space** domain using **Short**, **Medium**, and **Long** Videos, categorized by reasoning types, based on a subset of the dataset.

Difficulty	**Models**	**Size**	**Over. Avg.**	**Short Video Scenarios**				**Medium Video Scenarios**				**Long Video Scenarios**			
				Avg.	**Temporal**	**Spatial**	**Intent**	**Avg.**	**Temporal**	**Spatial**	**Intent**	**Avg.**	**Temporal**	**Spatial**	**Intent**
Hard	GPT 4o [18]	-	24.41	26.78	34.65	34.69	11.00	35.70	43.14	32.14	31.82	11.00	6.00	26.00	1.00
	Gemini 2.5 pro [13]	-	**29.76**	34.84	36.63	44.90	23.00	35.76	45.10	30.36	31.82	18.67	10.00	28.00	18.0
	Gemini 2.5 flash think [12]	-	28.67	32.13	35.64	37.75	23.00	35.20	37.25	41.07	27.27	18.67	6.00	36.00	14.00
	Gemini 2.5 flash no-think [12]	-	24.34	24.74	30.69	26.53	17.00	30.94	52.94	23.21	16.67	17.33	14.00	24.00	14.00
	Gemini 1.5 pro [8]	-	18.76	19.72	23.76	20.41	15.00	24.55	33.33	16.07	24.24	12.00	2.00	26.00	8.00
	Claude 3.5 [2]	-	28.71	33.76	35.64	31.63	34.00	28.87	37.26	35.71	13.63	16.00	12.00	26.00	10.0
	InternVL2.5 [7]	26B	23.78	21.33	26.00	31.00	7.00	32.00	46.00	32.00	18.00	18.00	16.00	24.00	14.00
	InternVL2.5 [7]	8B	22.67	20.00	18.00	33.00	9.00	30.00	46.00	30.00	14.00	18.00	16.00	28.00	10.00
	InternVL2.5 [7]	4B	19.56	18.67	18.00	28.00	8.00	28.00	34.00	24.00	26.00	12.00	8.00	22.00	6.00
	LLaVA Next [21]	32B	16.22	20.67	16.00	32.00	14.00	11.33	12.00	12.00	10.00	16.67	10.00	30.00	10.00
	LLaVA Video [48]	7B	19.78	19.33	12.00	35.00	11.00	24.67	26.00	30.00	18.00	15.33	10.00	28.00	8.00
	LLaVA OneVision [22]	7B	13.67	14.33	5.00	27.00	11.00	14.67	18.00	8.00	18.00	12.00	6.00	22.00	8.00
	Qwen2.5 VL [3]	32B	22.66	19.33	11.00	34.00	13.00	35.33	46.00	24.00	36.00	13.33	4.00	26.00	10.00
	Qwen2.5 VL [3]	7B	22.89	26.00	17.00	30.00	31.00	30.00	40.00	32.00	18.00	12.67	2.00	30.00	6.00
Medium	GPT 4o [18]	-	36.99	45.49	48.48	55.00	33.00	33.89	41.67	26.67	33.33	31.33	24.00	44.00	26.00
	Gemini 2.5 pro [13]	-	36.46	42.79	38.38	59.00	31.00	33.93	39.58	28.89	33.33	32.67	28.00	44.00	26.0
	Gemini 2.5 flash think [12]	-	**37.52**	47.82	46.47	56.00	41.00	36.99	43.75	42.22	25.00	28.00	12.00	44.00	28.00
	Gemini 2.5 flash no-think [12]	-	36.70	47.50	48.49	58.00	36.00	33.93	39.58	28.89	33.33	28.67	24.00	42.00	20.00
	Gemini 1.5 pro [8]	-	33.89	39.47	42.42	42.00	34.00	33.52	33.33	42.22	25	28.67	12.00	52.00	22.00
	Claude 3.5 [2]	-	35.35	41.78	35.35	50.00	40.00	35.60	39.58	42.22	25.00	28.67	16.00	44.00	26.0
	InternVL2.5 [7]	26B	35.11	36.00	39.00	50.00	19.00	36.67	50.00	36.00	24.00	32.67	30.00	40.00	28.00
	InternVL2.5 [7]	8B	34.66	37.33	43.00	57.00	12.00	35.33	42.00	46.00	18.00	31.33	26.00	44.00	24.00
	InternVL2.5 [7]	4B	33.89	39.67	38.00	53.00	28.00	32.67	44.00	28.00	26.00	29.33	16.00	46.00	26.00
	LLaVA Next [21]	32B	20.00	27.33	16.00	49.00	17.00	10.67	14.00	10.00	8.00	22.00	16.00	36.00	14.00
	LLaVA Video [48]	7B	25.67	25.00	20.00	34.00	26.00	28.67	36.00	28.00	22.00	23.33	14.00	40.00	16.00
	LLaVA OneVision [22]	7B	16.67	16.00	26.00	30.00	16.00	14.67	18.00	8.00	18.00	19.33	12.00	30.00	16.00
	Qwen2.5 VL [3]	32B	28.55	28.33	21.00	44.00	20.00	33.33	40.00	30.00	30.00	24.00	8.00	40.00	24.00
	Qwen2.5 VL [3]	7B	29.89	39.00	37.00	42.00	38.00	30.67	32.00	40.00	20.00	20.00	16.00	26.00	18.00
Easy	GPT 4o [18]	-	42.17	52.35	59.00	47.06	51.00	47.16	54.9	44.9	41.67	27.00	44.00	5.00	32.00
	Gemini 2.5 pro [13]	-	**54.56**	62.96	70.00	55.88	63.00	54.73	52.94	59.18	52.08	46.00	40.00	54.00	44.00
	Gemini 2.5 flash think [12]	-	50.00	67.56	69.00	65.69	68.00	44.45	52.94	40.82	39.58	38.00	32.00	38.00	44.00
	Gemini 2.5 flash no-think [12]	-	51.40	58.97	70.00	54.90	52.00	46.56	52.94	36.74	50.00	48.67	38.00	56.00	52.00
	Gemini 1.5 pro [8]	-	46.00	51.33	60.00	50.00	44.00	36.92	49.02	36.73	25.00	50.00	58.00	44.00	48.00
	Claude 3.5 [2]	-	48.59	60.33	61.00	50.00	70.00	36.35	35.29	51.02	22.73	49.33	64.00	44.00	40.0
	InternVL2.5 [7]	26B	52.55	61.00	62.00	59.00	62.00	45.33	58.00	44.00	34.00	**51.33**	62.00	62.00	30.00
	InternVL2.5 [7]	8B	50.11	55.67	55.00	60.00	52.00	44.67	58.00	42.00	34.00	50.00	54.00	64.00	32.00
	InternVL2.5 [7]	4B	44.89	53.33	46.00	60.00	54.00	37.33	48.00	38.00	26.00	44.00	44.00	48.00	40.00
	LLaVA Next [21]	32B	31.25	38.00	35.00	45.00	34.00	21.33	12.00	14.00	38.00	34.67	20.00	50.00	34.00
	LLaVA Video [48]	7B	31.44	33.00	30.00	31.00	38.00	33.33	38.00	36.00	26.00	28.00	16.00	32.00	36.00
	LLaVA OneVision [22]	7B	29.78	32.00	31.00	33.00	32.00	24.00	26.00	30.00	16.00	33.33	28.00	36.00	36.00
	Qwen2.5 VL [3]	32B	43.22	51.00	58.00	50.00	45.00	41.33	46.00	38.00	40.00	37.33	32.00	44.00	36.00
	Qwen2.5 VL [3]	7B	40.67	51.33	55.00	42.00	57.00	36.00	32.00	42.00	34.00	34.67	34.00	28.00	42.00

Land Space Analysis: As shown in Table 7, we present a comprehensive evaluation of model performance in the **Land Space** domain of M4R, categorized by reasoning type, video length, and difficulty level. In the easy setting, proprietary models such as Gemini 2.5 Pro and GPT-4o exhibit strong performance, achieving over 50% overall accuracy, with Gemini 2.5 Pro reaching the highest at 54.56%. Among open-source models, InternVL2.5 (26B) performs competitively with an overall accuracy of 52.55%. However, performance drops significantly across all models as task difficulty increases and video length extends. For instance, in hard tasks involving long videos, even the

Table 8: Evaluation of M4R in the **Air Space** domain using **Short**, **Medium**, and **Long** Videos, categorized by reasoning types, based on a subset of the dataset.

Difficulty	Models	Size	Over. Avg.	Short Video Scenarios				Medium Video Scenarios				Long Video Scenarios			
				Avg.	Temporal	Spatial	Intent	Avg.	Temporal	Spatial	Intent	Avg.	Temporal	Spatial	Intent
Hard	GPT 4o [18]	-	18.11	21.33	16.00	26.00	22.00	14.67	12.00	30.00	2.00	18.33	5.00	35.00	15.00
	Gemini 2.5 pro [13]	-	**31.39**	32.83	36.00	24.49	38.00	24.67	32.00	22.00	20.00	36.67	30.00	15.00	65.00
	Gemini 2.5 flash think [12]	-	25.78	26.00	26.00	18.00	34.00	21.33	28.00	18.00	18.00	30.00	30.00	10.00	50.00
	Gemini 2.5 flash no-think [12]	-	25.44	25.33	22.00	28.00	26.00	26.00	26.00	28.00	24.00	25.00	0.00	40.00	35.00
	Gemini 1.5 pro [8]	-	22.34	26.67	24.00	26.00	30.00	18.67	20.00	22.00	14.00	21.67	10.00	25.00	30.00
	Claude 3.5 [2]	-	24.22	26.00	18.00	32.00	28.00	23.33	20.00	28.00	22.00	23.33	10.00	40.00	20.0
	InternVL2.5 [7]	26B	17.33	19.33	24.00	26.00	10.00	19.33	16.00	32.00	10.00	13.33	10.00	10.00	20.00
	InternVL2.5 [7]	8B	18.22	18.67	20.00	28.00	8.00	19.33	16.00	30.00	12.00	16.67	5.00	35.00	10.00
	InternVL2.5 [7]	4B	15.33	15.33	14.00	10.00	22.00	14.00	16.00	18.00	8.00	16.67	15.00	30.00	5.00
	LLaVA Next [21]	32B	17.89	18.67	14.0	34.0	8.00	16.67	6.00	32.00	12.00	18.33	5.0	40.0	10.00
	LLaVA Video [48]	7B	14.78	16.67	14.00	28.00	8.00	12.67	6.00	22.00	10.00	15.00	5.00	30.00	10.00
	LLaVA OneVision [22]	7B	15.67	16.00	12.00	28.00	8.00	16.00	12.00	26.00	10.00	15.00	10.00	25.00	10.00
	Qwen2.5 VL [3]	32B	16.22	20.00	6.00	36.00	18.00	15.33	4.00	24.00	18.00	13.33	0.00	30.00	10.00
	Qwen2.5 VL [3]	7B	16.55	19.33	0.00	30.00	28.00	15.33	2.00	30.00	14.00	15.00	5.00	30.00	10.00
Medium	GPT 4o [18]	-	38.45	38.67	38.00	56.00	22.00	30.00	38.00	34.00	18.00	46.67	65.00	30.00	45.00
	Gemini 2.5 pro [13]	-	43.11	44.67	42.00	40.00	52.00	31.33	34.00	34.00	26.00	53.33	60.00	35.00	65.0
	Gemini 2.5 flash think [12]	-	39.78	39.33	32.00	38.00	48.00	30.00	34.00	28.00	28.00	50.00	65.00	15.00	70.00
	Gemini 2.5 flash no-think [12]	-	**49.67**	43.33	30.00	48.00	52.00	40.67	38.00	50.00	34.00	65.00	60.00	65.00	70.00
	Gemini 1.5 pro [8]	-	38.78	38.00	32.00	48.00	34.00	36.67	34.00	52.00	24.00	41.67	30.00	55.00	40.00
	Claude 3.5 [2]	-	39.67	38.00	26.00	40.00	48.00	36.00	32.00	54.00	22.00	45.00	50.00	35.00	50.0
	InternVL2.5 [7]	26B	28.67	31.33	28.00	58.00	8.00	24.67	12.00	50.00	12.00	30.00	25.00	45.00	20.00
	InternVL2.5 [7]	8B	34.33	30.00	20.00	58.00	12.00	34.67	32.00	50.00	22.00	38.33	40.00	45.00	30.00
	InternVL2.5 [7]	4B	32.22	29.33	28.00	44.00	16.00	34.00	30.00	54.00	18.00	33.33	35.00	40.00	25.00
	LLaVA Next [21]	32B	26.11	24.67	18.0	40.0	16.00	25.33	18.0	40.0	18.00	28.33	25.0	40.0	20.00
	LLaVA Video [48]	7B	24.00	25.33	24.00	36.00	16.00	20.00	16.00	26.00	18.00	26.67	15.00	45.00	20.00
	LLaVA OneVision [22]	7B	23.67	23.33	20.00	34.00	16.00	22.67	20.00	32.00	16.00	25.00	20.00	35.00	20.00
	Qwen2.5 VL [3]	32B	33.34	32.67	12.00	48.00	38.00	30.67	22.00	50.00	20.00	36.67	20.00	60.00	30.00
	Qwen2.5 VL [3]	7B	28.00	24.67	16.00	24.00	34.00	26.00	24.00	26.00	28.00	33.33	35.00	20.00	45.00
Easy	GPT 4o [18]	-	40.67	35.33	30.00	28.00	48.00	36.67	24.00	38.00	48.00	50.00	45.00	50.00	55.00
	Gemini 2.5 pro [13]	-	**52.56**	56.00	60.00	48.00	60.00	40.00	40.00	36.00	44.00	61.67	75.00	35.00	75.0
	Gemini 2.5 flash think [12]	-	50.67	49.33	40.00	46.00	62.00	46.00	46.00	44.00	48.00	56.67	55.00	40.00	75.00
	Gemini 2.5 flash no-think [12]	-	50.78	49.33	36.00	52.00	60.00	48.00	40.00	50.00	54.00	55.00	60.00	50.00	55.00
	Gemini 1.5 pro [8]	-	43.00	45.33	36.00	44.00	56.00	42.00	48.00	32.00	46.00	41.67	35.00	50.00	40.00
	Claude 3.5 [2]	-	42.45	38.00	34.00	38.00	42.00	42.67	30.00	56.00	42.00	46.67	40.00	45.00	55.0
	InternVL2.5 [7]	26B	36.11	35.33	36.00	44.00	26.00	34.67	28.00	46.00	30.00	38.33	30.00	40.00	45.00
	InternVL2.5 [7]	8B	38.44	36.67	28.00	46.00	36.00	35.33	32.00	42.00	32.00	43.33	60.00	40.00	30.00
	InternVL2.5 [7]	4B	40.33	43.33	42.00	50.00	38.00	39.33	30.00	44.00	44.00	38.33	35.00	60.00	20.00
	LLaVA Next [21]	32B	33.22	36.67	36.00	42.00	32.00	31.33	36.00	32.00	26.00	31.67	35.00	30.00	30.00
	LLaVA Video [48]	7B	33.22	33.33	34.00	38.00	28.00	34.67	34.00	38.00	32.00	31.67	35.00	30.00	30.00
	LLaVA OneVision [22]	7B	33.22	33.33	34.00	38.00	28.00	34.67	34.00	38.00	32.00	31.67	35.00	30.00	30.00
	Qwen2.5 VL [3]	32B	52.45	50.00	34.00	56.00	60.00	50.67	40.00	54.00	58.00	56.67	55.00	60.00	55.00
	Qwen2.5 VL [3]	7B	39.89	33.33	28.00	18.00	54.00	38.00	48.00	16.00	50.00	48.33	55.00	30.00	60.00

best-performing models fall below 30% average accuracy. These trends highlight the limitations of current multimodal models in handling complex, real-world reasoning—particularly for extended temporal sequences, fine-grained spatial relations, and intent inference. While models like GPT-4o and Gemini 1.5 Pro show relatively strong performance on medium-difficulty tasks (36.99% and 33.89%, respectively), the results underscore the persistent challenges in achieving robust reasoning across diverse open-space scenarios.

Air Space Analysis: Table 8 presents model performance in the **Air Space** domain of M4R, evaluated across short, medium, and long video scenarios, and categorized by temporal, spatial, and intent reasoning tasks. In the easy setting, **Gemini 2.5 Pro** achieves the highest overall accuracy (52.56%), outperforming all other models, including GPT-4o and Qwen2.5 (32B). In the medium setting, Gemini 2.5 Pro again leads with 43.11%, followed closely by Gemini 1.5 Pro (38.78%) and GPT-4o (38.45%). For hard tasks, which are the most challenging, **Gemini 2.5 Pro** remains the top performer with 31.39%. These results highlight the ability of the Gemini family of models to maintain performance in complex, dynamic airspace environments. Meanwhile, GPT-4o and Qwen2.5 models show competitive results in easier tasks but exhibit notable drops as the reasoning complexity increases, revealing current limitations in handling temporal, spatial, and intent-based challenges in aerial domains. Moreover, Table 9 presents model performance on the M4R benchmark in the **Water Space** domain, covering both river and ocean scenarios across varying reasoning types and difficulty levels. Gemini models consistently outperform other models across all settings.

These findings demonstrate M4R's ability to *reveal the limitations* of existing multimodal models, particularly in safety-critical and physically grounded domains. By highlighting domain-specific reasoning gaps, especially in underexplored high-stakes environments such as autonomous driving, ship navigation, and airspace, M4R serves as a tool for guiding the development of more robust, temporally aware, and intent-aware multimodal systems.

Table 9: Evaluation of M4R in the **Water Space** domain using **River** and **Ocean** Videos, categorized by reasoning types, based on a subset of the dataset.

Difficulty	Models	Size	Over. Avg.	River Scenarios				Ocean Scenarios			
				Avg.	Temporal	Spatial	Intent	Avg.	Temporal	Spatial	Intent
	GPT4o [18]	-	22.10	28.20	38.46	26.92	19.23	16.00	18.00	18.00	12.00
	Gemini 2.5 pro [13]	-	**29.64**	34.62	23.08	34.62	46.15	24.67	38.00	16.00	20.00
	Gemini 2.5 flash think [12]	-	27.36	32.05	30.77	26.92	38.46	22.67	30.00	22.00	16.00
	Gemini 2.5 flash no-think [12]	-	27.44	28.21	42.31	19.23	23.08	26.67	36.00	20.00	24.00
	Gemini 1.5 pro [8]	-	26.02	26.92	23.08	30.77	26.92	25.11	34.00	20.93	20.41
	Claude 3.5 [2]	-	25.44	28.20	19.23	19.23	46.15	22.67	26.00	22.00	20.00
Hard	InternVL2.5 [7]	26B	22.54	23.08	15.38	19.23	34.62	22.00	18.00	28.00	20.00
	InternVL2.5 [7]	8B	21.90	21.79	7.69	26.92	30.77	22.00	16.00	28.00	22.00
	InternVL2.5 [7]	4B	20.92	20.51	19.23	19.23	23.08	21.33	16.00	26.00	22.00
	LLaVA Next [21]	32B	14.39	11.54	7.69	19.23	7.69	15.33	8.00	30.00	8.00
	LLaVA Video [48]	7B	14.00	16.67	15.38	23.08	11.54	11.33	8.00	20.00	6.00
	LLaVA OneVision [22]	7B	15.67	16.67	11.54	26.92	11.54	14.67	8.00	28.00	8.00
	Qwen2.5 VL [3]	32B	13.39	14.10	7.69	23.08	11.54	12.67	8.0	24.0	6.00
	Qwen2.5 VL [3]	7B	14.67	16.67	7.69	30.77	11.54	12.67	6.00	24.00	8.00
	GPT 4o [18]	-	38.49	42.31	50.00	53.85	23.08	34.67	36.00	48.00	20.00
	Gemini 2.5 pro [13]	-	41.77	44.87	30.77	61.54	42.31	38.67	48.00	46.00	22.00
	Gemini 2.5 flash think [12]	-	**48.26**	53.85	61.54	57.70	42.31	42.67	52.00	42.00	34.00
	Gemini 2.5 flash no-think [12]	-	46.12	50.00	46.15	57.69	46.15	42.00	56.00	44.00	26.00
	Gemini 1.5 pro [8]	-	46.31	53.84	46.15	65.38	50.00	38.78	34.00	49.02	33.33
	Claude 3.5 [2]	-	38.62	35.90	34.62	50.00	23.08	41.33	42.00	54.00	28.00
Medium	InternVL2.5 [7]	26B	41.77	44.87	30.77	57.69	46.15	38.67	24.00	62.00	30.00
	InternVL2.5 [7]	8B	41.08	46.15	34.62	61.54	42.31	36.00	34.00	60.00	14.00
	InternVL2.5 [7]	4B	44.36	48.72	23.08	65.38	57.69	40.00	28.00	60.00	32.00
	LLaVA Next [21]	32B	20.88	23.08	11.54	38.46	19.23	18.67	10.00	30.00	16.00
	LLaVA Video [48]	7B	21.92	20.51	19.23	26.92	15.38	23.33	20.00	30.00	20.00
	LLaVA OneVision [22]	7B	22.54	23.08	19.23	30.77	19.23	22.00	14.00	34.00	18.00
	Qwen2.5 VL [3]	32B	33.31	34.62	19.23	50.00	34.62	32.00	20.00	50.00	26.00
	Qwen2.5 VL [3]	7B	24.08	29.49	19.23	30.77	38.46	18.67	18.00	26.00	12.00
	GPT 4o [18]	-	50.51	57.69	57.69	50.00	65.38	43.33	66.00	34.00	30.00
	Gemini 2.5 pro [13]	-	61.05	64.10	57.69	57.69	76.92	58.00	72.00	50.00	52.00
	Gemini 2.5 flash think [12]	-	**62.03**	65.39	80.77	42.31	73.08	58.67	70.00	52.00	54.00
	Gemini 2.5 flash no-think [12]	-	58.18	57.69	57.69	38.46	76.92	58.67	80.00	42.00	54.00
	Gemini 1.5 pro [8]	-	50.69	52.56	42.31	61.54	53.85	48.81	50.00	46.43	50.00
	Claude 3.5 [2]	-	49.39	47.44	50.00	53.85	38.46	51.33	62.00	52.00	40.00
Easy	InternVL2.5 [7]	26B	55.05	64.10	65.38	57.69	69.23	46.00	50.00	50.00	38.00
	InternVL2.5 [7]	8B	53.47	60.26	69.23	46.15	65.38	46.67	46.00	54.00	40.00
	InternVL2.5 [7]	4B	53.87	56.41	53.85	57.69	57.69	51.33	52.00	56.00	46.00
	LLaVA Next [21]	32B	35.59	37.18	26.92	53.85	30.77	34.00	30.00	38.00	34.00
	LLaVA Video [48]	7B	31.03	32.05	30.77	34.62	30.77	30.00	22.00	38.00	30.00
	LLaVA OneVision [22]	7B	33.00	33.33	34.62	34.62	30.77	32.67	28.00	38.00	32.00
	Qwen2.5 VL [3]	32B	52.77	61.54	53.85	61.54	69.23	44.00	40.00	54.00	38.00
	Qwen2.5 VL [3]	7B	31.31	34.62	38.46	19.23	46.15	28.00	36.00	22.00	26.00

4.4 Model Error Analysis

To demonstrate the effectiveness of our benchmark and evaluate the performance of state-of-the-art (SOTA) models, we conduct a qualitative analysis of model predictions on the M4R benchmark. As shown in Figure 5, the analysis highlights persistent challenges in spatial, temporal, and intent reasoning across open-space environments, particularly in land and air domains. Despite the strong overall performance of leading multimodal models such as ChatGPT-4o and Gemini 2.5, the results reveal consistent failure cases in real-world scenarios. For example, both models struggle with accurately identifying spatial relationships (e.g., relative positions of vehicles), counting dynamic objects over time (e.g., cars in motion), and understanding goal-directed interactions (e.g., airplane passing events).

These failure cases underscore the limitations of current models in handling safety-critical, perception-intensive tasks. By providing richly annotated, video-based tasks that demand multi-step reasoning grounded in physics, causality, and spatial understanding, M4R serves as a rigorous diagnostic benchmark. Our findings highlight the necessity of such benchmarks for advancing the robustness, safety, and real-world applicability of large multimodal systems.

4.5 Ablation Experiments

In our experiments, due to the high cost of evaluating all data points, we adopt a uniform sampling strategy to select a representative subset of tasks. Specifically, for each reasoning type, we sample 50 tasks when the total number of available tasks is fewer than 500, and 100 tasks when the number

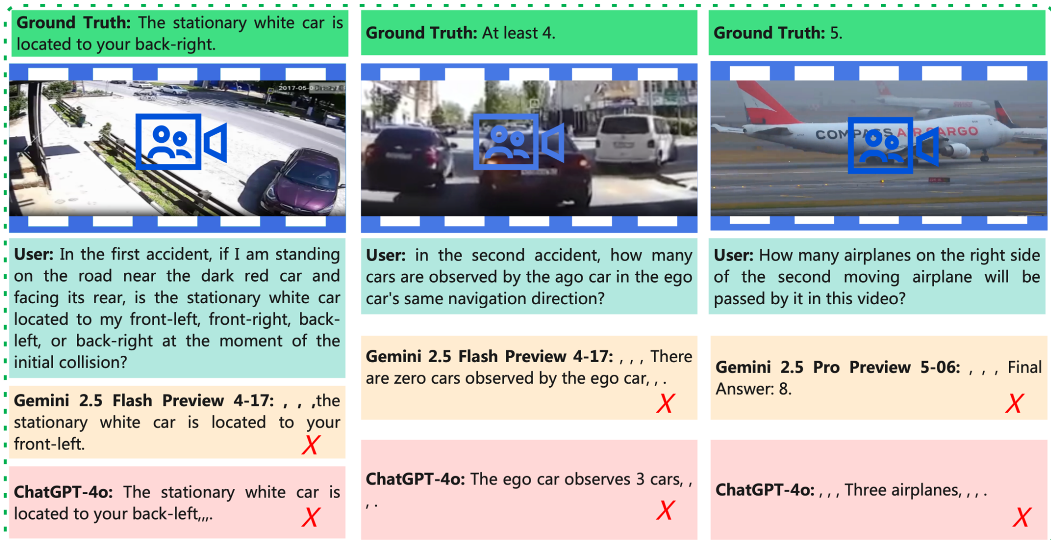


Figure 5: Qualitative error analysis of state-of-the-art multimodal models (Gemini 2.5 and ChatGPT-4o) on the M4R benchmark. Each example illustrates a failure case in a different reasoning category: spatial reasoning (left), temporal reasoning (middle), and intent reasoning (right). Despite their capabilities, both models struggle with spatial localization, counting dynamic objects, and understanding goal-directed motion in real-world open-space scenarios.

exceeds 500. The M4R spans three open-space scenarios—*land space*, *air space*, and *water space*—each with three video lengths (short, medium, long), three difficulty levels (easy, medium, hard), and three reasoning types: temporal, spatial, and intent-based reasoning.

Following this sampling strategy, we evaluate a total of 3,798 tasks, evenly distributed across the three reasoning types: 1,266 *spatial reasoning*, 1,266 *temporal-causal reasoning*, and 1,266 *intent and goal reasoning* tasks.

To assess the reliability of this sampling approach, we conduct an ablation study comparing model performance on sampled tasks versus the full set of data points in the **land space (short, easy)** setting. We use InternVL 2.5, one of the leading open-source multimodal models, which ranks highly on several leaderboards such as [6] and [7]. As shown in Table 10, performance on the sampled subset is comparable to, and in some cases slightly better than, performance on the full dataset. These results validate the effectiveness of our sampling strategy in preserving benchmark consistency while reducing evaluation cost.

Table 10: Performance Comparison on **Land Space Short** (Easy): Full vs. Sample Data Points

Model	Full Data Points				Sample Data Points			
	Avg.	Temporal	Spatial	Intent	Avg.	Temporal	Spatial	Intent
InternVL2_5-26B	55.62	57.61	50.37	58.88	61.00	62.00	59.00	62.00
InternVL2_5-8B	49.26	51.89	48.57	47.31	55.67	55.00	60.00	52.00
InternVL2_5-4B	50.65	50.17	50.70	51.10	55.33	52.00	55.00	59.00

5 Conclusion

In this work, we introduce M4R, a large-scale benchmark for evaluating multimodal understanding and reasoning in real-world open-space environments. Spanning three critical domains—land, air, and water, M4R provides richly annotated, video-based tasks designed to assess model performance across three fundamental reasoning dimensions: temporal reasoning, spatial reasoning, and intent and goal inference. The benchmark encompasses a broad range of scenarios, video lengths, and difficulty levels, enabling comprehensive evaluation in safety-critical, perception-intensive settings.

[6] https://enxinsong.com/Video-MMLU-web/
[7] https://huggingface.co/spaces/opencompass/open_vlm_leaderboard

Through extensive qualitative and quantitative analyses, we demonstrate that even state-of-the-art multimodal models, both proprietary systems such as ChatGPT-4o and Gemini 2.5, and leading open-source models like Qwen and InternVL, exhibit significant limitations when reasoning over complex, dynamic physical environments. These results underscore the need for more robust, temporally-aware, and goal-sensitive multimodal systems capable of reliable understanding in real-world scenarios. We hope that M4R will serve as a valuable resource for the research community and help advance the development of safer, more generalizable, and practically deployable multimodal AI systems.

Acknowledgment

We thank NVIDIA for generously providing the computational resources used in our experiments.

References

[1] Guillermo Franco Abellán, Matteo Braglia, Mario Ballardini, Fabio Finelli, and Vivian Poulin. Probing early modification of gravity with planck, act and spt. *Journal of Cosmology and Astroparticle Physics*, 2023(12):017, 2023.

[2] Anthropic. Introducing claude 3.5 sonnet. *https://www.anthropic.com/news/claude-3-5-sonnet*, 2024.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[4] Wentao Bao, Qi Yu, and Yu Kong. Uncertainty-based traffic accident anticipation with spatio-temporal relational learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2682–2690, 2020.

[5] Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark. *arXiv preprint arXiv:2410.03051*, 2024.

[6] Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Frank Wang. Rex-time: A benchmark suite for reasoning-across-time in videos. *Advances in Neural Information Processing Systems*, 37:28662–28673, 2024.

[7] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[8] Google DeepMind. Gemini 1.5 technical report. `https://deepmind.google/technologies/gemini/#gemini-15`, 2024. Accessed: 2025-05-12.

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

[10] Hanan Gani, Rohit Bharadwaj, Muzammal Naseer, Fahad Shahbaz Khan, and Salman Khan. Vane-bench: Video anomaly evaluation benchmark for conversational lmms. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3123–3140, 2025.

[11] Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. Figstep: Jailbreaking large vision-language models via typographic visual prompts. *arXiv preprint arXiv:2311.05608*, 2023.

[12] Google. Gemini 2.5 flash. `https://storage.googleapis.com/model-cards/documents/gemini-2.5-flash.pdf`, July 2025. Accessed: 2025-06-26.

[13] Google. Gemini 2.5: Our most intelligent ai model. *https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/gemini-2-5-thinking*, 2025.

[14] Yu Guo, Ryan Wen Liu, Jingxiang Qu, Yuxu Lu, Fenghua Zhu, and Yisheng Lv. Asynchronous trajectory matching-based multimodal maritime data fusion for vessel traffic surveillance in inland waterways. *IEEE Transactions on Intelligent Transportation Systems*, 24(11):12779–12792, 2023.

[15] Xuehai He, Weixi Feng, Kaizhi Zheng, Yujie Lu, Wanrong Zhu, Jiachen Li, Yue Fan, Jianfeng Wang, Linjie Li, Zhengyuan Yang, et al. Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos. *arXiv preprint arXiv:2406.08407*, 2024.

[16] Xuehai He, Weixi Feng, Kaizhi Zheng, Yujie Lu, Wanrong Zhu, Jiachen Li, Yue Fan, Jianfeng Wang, Linjie Li, Zhengyuan Yang, et al. Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos. In *The Thirteenth International Conference on Learning Representations*, 2025.

[17] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.

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

[19] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025.

[20] Minkuk Kim, Hyeon Bae Kim, Jinyoung Moon, Jinwoo Choi, and Seong Tae Kim. Hicm^2: Hierarchical compact memory modeling for dense video captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4293–4301, 2025.

[21] Bo Li, Hao Zhang, Kaichen Zhang, Dong Guo, Yuanhan Zhang, Renrui Zhang, Feng Li, Ziwei Liu, and Chunyuan Li. Llava-next: What else influences visual instruction tuning beyond data?, May 2024.

[22] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

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

[24] Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *European Conference on Computer Vision*, pages 386–403. Springer, 2024.

[25] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024.

[26] Weidi Luo, Siyuan Ma, Xiaogeng Liu, Xiaoyu Guo, and Chaowei Xiao. Jailbreakv: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. *arXiv preprint arXiv:2404.03027*, 2024.

[27] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

[28] Dilip K Prasad, Deepu Rajan, Lily Rachmawati, Eshan Rajabally, and Chai Quek. Video processing from electro-optical sensors for object detection and tracking in a maritime environment: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 18(8):1993–2016, 2017.

[29] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 21527–21536, 2024.

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

[31] Ankit Parag Shah, Jean-Bapstite Lamare, Tuan Nguyen-Anh, and Alexander Hauptmann. Cadp: A novel dataset for cctv traffic camera based accident analysis. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–9. IEEE, 2018.

[32] Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. *arXiv preprint arXiv:2307.14539*, 2023.

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

[34] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European Conference on Computer Vision*, pages 256–274, 2024.

[35] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024.

[36] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. Moviechat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*, 2023.

[37] Enxin Song, Wenhao Chai, Weili Xu, Jianwen Xie, Yuxuan Liu, and Gaoang Wang. Video-mmlu: A massive multi-discipline lecture understanding benchmark. *arXiv preprint arXiv:2504.14693*, 2025.

[38] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, XianPeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. In *8th Annual Conference on Robot Learning*, 2025.

[39] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024.

[40] Hongchen Wei, Zhihong Tan, Yaosi Hu, Chang Wen Chen, and Zhenzhong Chen. Longcaptioning: Unlocking the power of long video caption generation in large multimodal models. *arXiv preprint arXiv:2502.15393*, 2025.

[41] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

[42] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021.

[43] Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are vlms ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives. *arXiv preprint arXiv:2501.04003*, 2025.

[44] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.

[45] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171*, 2024.

[46] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

[47] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024.

[48] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

[49] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024.

[50] Kaiwen Zhou, Chengzhi Liu, Xuandong Zhao, Anderson Compalas, Dawn Song, and Xin Eric Wang. Multimodal situational safety. *arXiv preprint arXiv:2410.06172*, 2024.

[51] Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *arXiv preprint arXiv:2411.00836*, 2024.

Appendix

A Limitation and Impact

Limitation Our benchmark provides a valuable tool for evaluating model performance in open-space environments. However, due to the large scale of the dataset, evaluating all data points is computationally expensive. As a result, we were unable to perform large-scale testing with high-cost proprietary models such as ChatGPT and Gemini. In future work, we plan to explore more efficient evaluation strategies and extend our analysis to a broader set of models, including closed-source systems.

Impact This benchmark offers a new direction for advancing multi-modal model development in open-space, safety-critical, and physically grounded real-world environments. By emphasizing temporal, spatial, and intent-based reasoning in diverse video scenarios, this benchmark can be useful to guide the design of more robust and reliable multi-modal systems. While this research seeks to advance the capabilities of AI in complex settings, we do not identify any specific societal risks or consequences requiring special attention at this time.

B Annotation and Detailed Examples

During data annotation, we first define the question types, then watch each video to design corresponding questions and annotate the answers. Our dataset contains approximately 2101 videos and 19,136 question–answer pairs, evenly distributed across three difficulty levels: easy (1/3), medium (1/3), and hard (1/3). The difficulty is determined by both the number and type of answer choices. Hard questions typically include 12 choices for temporal and intent reasoning, and 4 for spatial reasoning, requiring precise selection. Medium questions generally offer 6 choices for temporal and intent reasoning, and 3 for spatial reasoning, often involving interval-based options. Easy questions usually present 3 choices, or 2 for spatial reasoning, and also rely on interval-based distinctions.

Moreover, we provide several example scenarios illustrating understanding and reasoning in open space, as shown in Figure 6. Moreover, as illustrated in Figure 7, we present a detailed question-and-answer example. For each open-space reasoning setting, we include three video lengths, short, medium, and long, each featuring tasks designed to evaluate temporal, spatial, and intent reasoning.

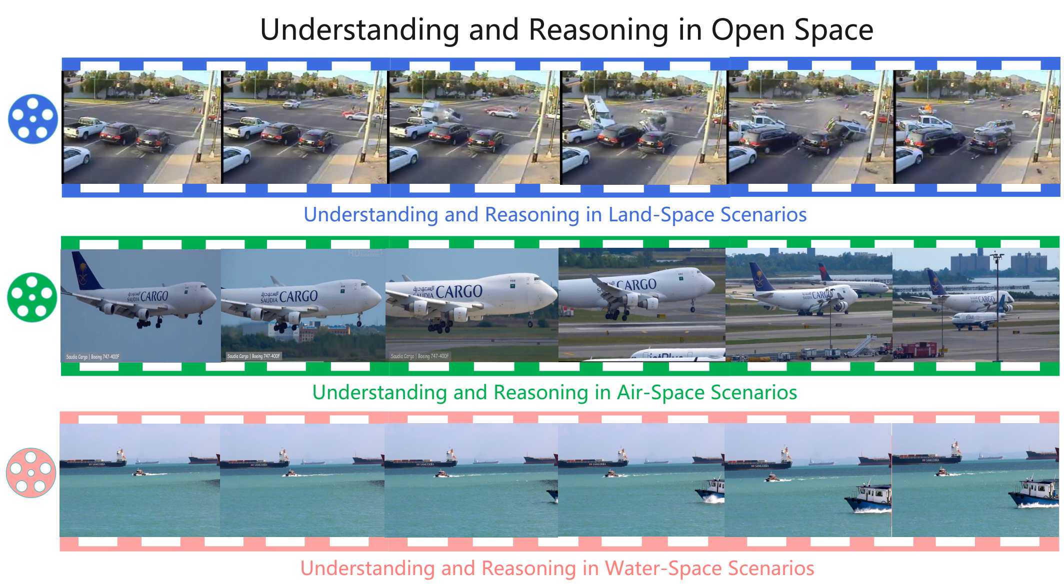


Figure 6: Example Scenarios of Understanding and Reasoning in Open Space

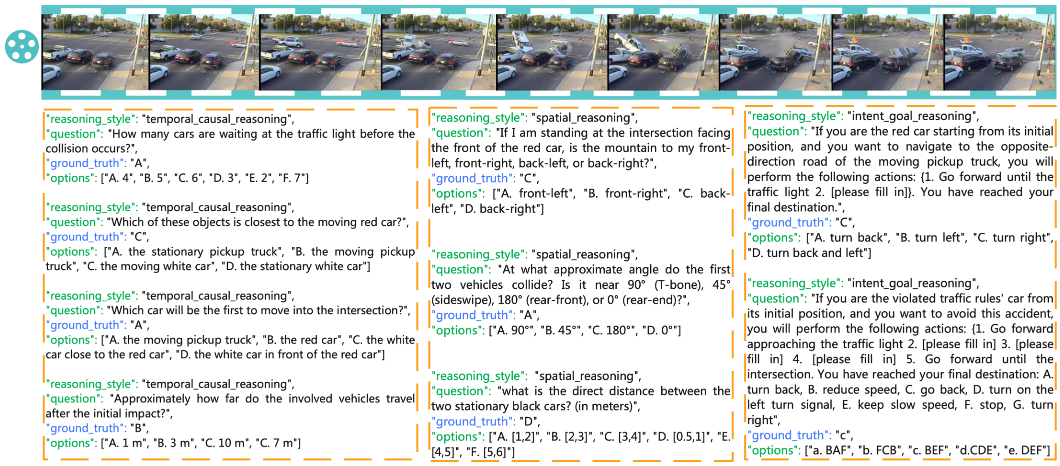


Figure 7: A question and answer example: For each open-space reasoning setting, we include three types of video lengths: short, medium, and long. Each video length includes tasks designed to evaluate temporal reasoning, spatial reasoning, and intent reasoning.

C Detailed Experiment Settings

In our experiments, we build upon the `lmms-eval` framework [47] as the foundation for our benchmark and extend it to support the specific requirements of M4R. All experiments with open-source models were conducted on a Linux system equipped with 8 × NVIDIA A100 GPUs, and experiments with closed-source models were run on a single NVIDIA A100 GPU. Key hyperparameters used for model evaluation are summarized in Table 11. More detailed experimental settings are available in our code: `https://open-space-reasoning.github.io`.

Table 11: Key parameters used in the experiments.

Parameters	value	Parameters	value
sample size	1	number of processes	8
max pixels (Qwen 2.5)	12845056	use-flash-attention-2 (Qwen 2.5)	False
interleave visuals (Qwen 2.5)	True	enable-chunked-prefill (InternVL 2.5)	True
gpu-memory-utilization (InternVL 2.5)	0.6	max-num-seqs (InternVL 2.5)	1
conv-template (LLava-Video)	qwen-1-5	video-decode-backend (LLava-Video)	record
max-frames-num (LLava-Video)	22	mm-spatial-pool-mode (LLava-Video)	average
mm-newline-position (LLava-Video)	grid	mm-resampler-location (LLava-Video)	after
conv-template (llava-onevision)	qwen-1-5	device-map (llava-onevision)	auto
model-name (llava-onevision)	llava-qwen		