

STI-Bench: Are MLLMs Ready for Precise Spatial-Temporal World Understanding?

Yun Li^{1,2}, Yiming Zhang^{1,3}, Tao Lin¹, XiangRui Liu^{1,4}, Wenxiao Cai⁵, Zheng Liu⁴, Bo Zhao¹

¹School of AI, Shanghai Jiao Tong University

²China University of Geosciences, ³Nanyang Technological University, ⁴BAAI, ⁵Stanford University

Corresponding to <bo.zhao@sjtu.edu.cn>

Abstract

The use of Multimodal Large Language Models (MLLMs) as an end-to-end solution for Embodied AI and Autonomous Driving has become a prevailing trend. While MLLMs have been extensively studied for visual semantic understanding tasks, their ability to perform precise and quantitative spatial-temporal understanding in real-world applications remains largely unexamined, leading to uncertain prospects. To evaluate models' Spatial-Temporal Intelligence, we introduce STI-Bench, a benchmark designed to evaluate MLLMs' spatial-temporal understanding through challenging tasks such as estimating and predicting the appearance, pose, displacement, and motion of objects. Our benchmark encompasses a wide range of robot and vehicle operations across outdoor, indoor, and desktop scenarios. The extensive experiments reveal that the state-of-the-art MLLMs still struggle in real-world spatial-temporal understanding, especially in tasks requiring precise distance estimation and motion analysis. Paper Page: <https://mint-sjtu.github.io/STI-Bench.io/>

1. Introduction

Recent advances in Multimodal Large Language Models (MLLMs) [1, 4, 12, 29, 35, 37–39, 44, 50] have positioned them as powerful tools for numerous vision and multimodal tasks. MLLMs achieve impressive performance in general Visual Question Answering tasks [3], which mainly focus on 2D visual perception and semantic question answering [18, 19, 21, 25, 40, 41, 52]. Beyond 2D visual perception, MLLMs are increasingly employed as end-to-end solutions for Embodied AI [7–9, 15, 23, 26, 32, 46] and Autonomous Driving [22, 42, 43, 48]. Such tasks require MLLMs to understand 3D space and time, and predict optimal manipulation strategies for robotic and vehicular systems. Despite these explorations, the question remains: Are MLLMs ready for precise spatial-temporal world un-

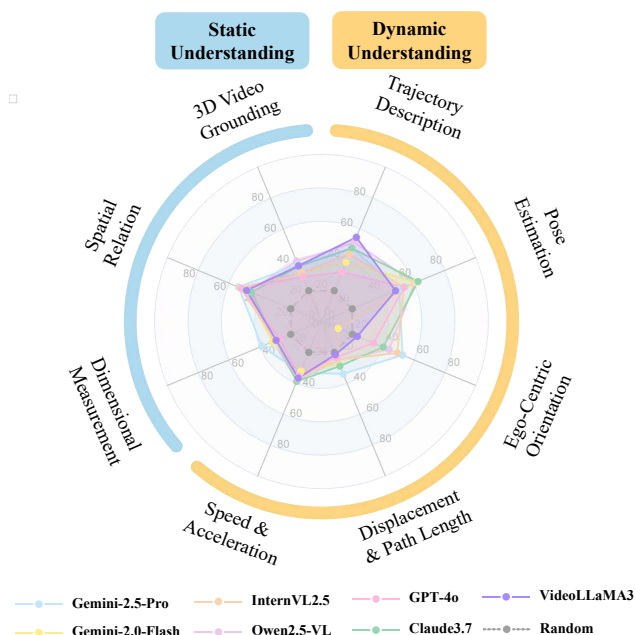


Figure 1. We evaluate state-of-the-art MLLMs on STI-Bench for precise and quantitative spatial-temporal understanding using video inputs. Results indicate the significant challenge in all tasks.

derstanding?

To answer this question, we propose a **Spatial-Temporal Intelligence Benchmark (STI-Bench)**, designed to evaluate MLLMs' spatial-temporal world understanding capability. We evaluate MLLMs using single video or multiple images as input instead of 3D point clouds. The main reasons are: 1) the majority of state-of-the-art models, e.g., GPT-4o [33] and Gemini [38], accept images or video as input rather than 3D point clouds; 2) Videos are more frequently used in daily life and contain sufficient information to infer the spatial-temporal environment.

STI-Bench contains 300 videos and over 2,000 QA pairs,

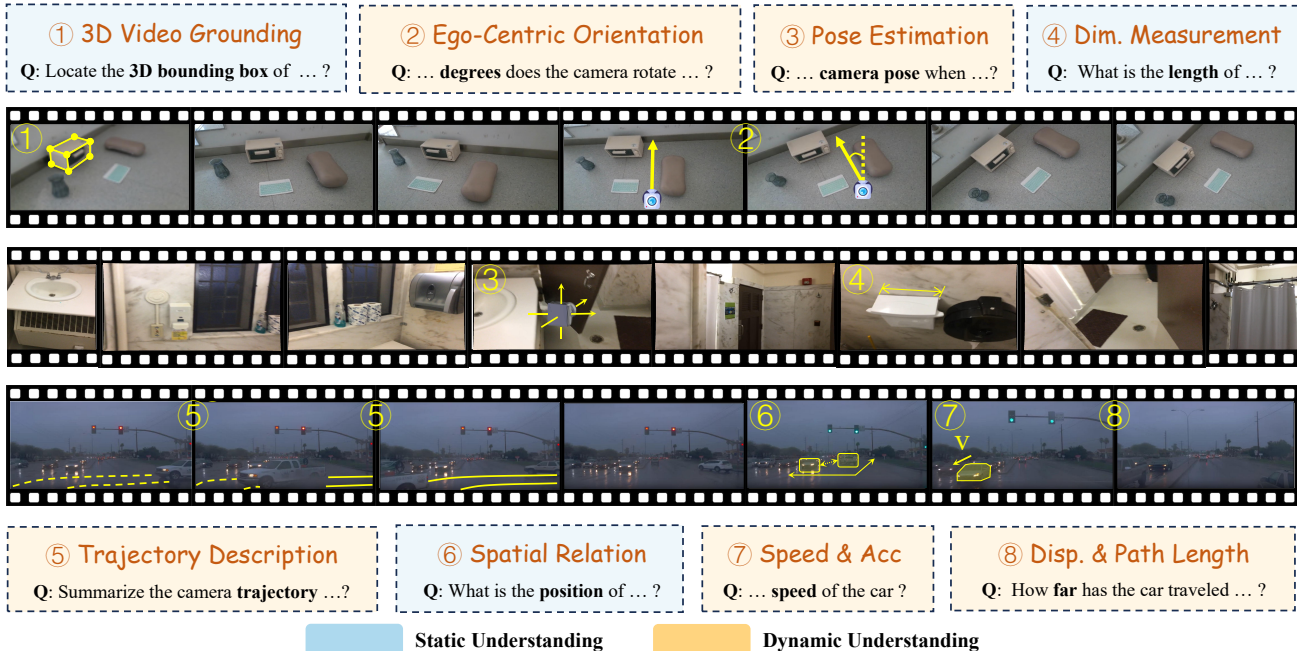


Figure 2. **Overview of STI-Bench.** We selected the most representative videos from each dataset scene and provided a few simple questions for demonstration.

covering three major scenarios: *Desktop*, *Indoor*, and *Outdoor*. The videos are sourced from Waymo [36], ScanNet [16], and Omni6DPose [51] respectively, encompassing a broad spectrum of real-world environments. As illustrated in Figure 2, we design eight distinct tasks to evaluate models’ ability in static spatial measurement and grounding, and dynamic tasks including speed, acceleration and trajectory estimation.

Through extensive experiments as illustrated in Figure 1, we observe that even the most advanced MLLMs struggle with real-world spatial-temporal understanding, especially in tasks requiring precise distance estimation and motion analysis. Our error analysis reveals three fundamental limitations: inaccurate spatial quantification, flawed temporal dynamics understanding, and weak cross-modal grounding and integration. These insights highlight the significant challenges MLLMs face in precisely understanding spatial-temporal information from videos. We believe STI-Bench will serve as an important touchstone that guides the community to develop better MLLMs for Embodied AI, Autonomous Driving tasks and beyond.

In summary, our main contributions include:

- We present STI-Bench, comprising over 300 videos and over 2,000 tailored questions across outdoor, indoor, and desktop scenarios, providing a systematic quantitative assessment of MLLMs’ spatial-temporal understanding capabilities.

- We conduct an in-depth study of state-of-the-art video-based MLLMs on STI-Bench, identify key error patterns in spatial-temporal reasoning, and provide empirical insights that can help the community develop more reliable MLLMs for embodied applications.

2. Related Work

2.1. Multimodal Large Language Models

MLLMs have achieved groundbreaking performance in visual understanding [1, 4, 12, 38], leveraging large language models (LLMs) [37, 39, 44] and visual encoders. Recent advancements have extended multimodal learning to video understanding. Classical works include VideoChat[24], which enables interactive video-based dialogue by integrating multimodal understanding. Subsequent models like Video-LLaVA[30] enhance visual-language alignment through large-scale vision-language pretraining, extending LLaVA[31]’s capability to process video inputs effectively. Recent works like Qwen2.5-VL [44] excel in long-video understanding and temporal localization by incorporating absolute temporal encoding, enabling models to capture relationships among video frames more effectively.

2.2. Spatial Understanding with MLLMs

Video MLLMs have focused heavily on semantic understanding. However, spatial understanding remains a significant challenge, inspiring recent contributions [9, 11, 14].

Benchmark	QA Pairs	Data	Env.	Scene			View		Evaluation		Spatial-Temporal			
				D	I	O	Ego	Allo.	Num.	Desc.	Dist.	Dir.	Vel.	Traj.
SAT [34]	218k	I	S	×	✓	×	✓	✓	✓	✓	×	×	×	×
VSI-Bench [45]	5,156	V	R	✓	×	×	✓	✓	✓	✓	✓	×	×	✓
EmbSpatial-Bench [17]	3,640	I	R	×	✓	×	✓	×	×	✓	×	×	×	×
EmbodiedAgentInterface [26]	448	-	S	×	✓	×	✓	×	-	-	×	×	×	×
EmbodiedEval [15]	328	I/V	S	×	✓	✓	✓	×	-	-	×	×	×	×
EmbodiedBench [46]	1,128	I	S	×	✓	✓	✓	×	-	-	×	×	×	×
WorldSense [6]	3,172	V	R	✓	✓	✓	✓	✓	×	✓	×	×	×	×
MLVU [52]	3,102	V	R	✓	✓	✓	✓	✓	×	✓	×	×	×	×
Video-MMMU [21]	300	V	S	×	×	×	×	×	×	✓	×	×	×	×
STI-Bench	2,064	V	R	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1. **Comparison of STI-Bench with existing benchmarks.** **Data** represents the source of our QA data, where **V** stands for Video and **I** stands for Image. **Env.** indicates the environment in which the data is generated, where **S** represents Simulation and **R** represents Real. The two columns under **View** indicate whether the dataset includes Ego-centric and Allocentric perspectives. The two columns under **Evaluation** specify whether the ground truth is presented in numerical or textual form. The four columns under **Spatial-Temporal** indicate whether the benchmark evaluates spatial distance, direction (with angular precision), velocity, or a precise and comprehensive trajectory description.

This progress represents a significant step toward developing world models and embodied agents. Recent advancements in embodied intelligence have explored integrating large-scale MLLMs into robotic control, enabling better generalization and semantic reasoning. RT-2 [8] introduces a vision-language-action framework that transfers web-scale knowledge to robotic control by representing actions as tokens alongside visual and language data. Building on this idea, GR-2 [10] extends generalist robot control across diverse embodiments using a Transformer-based architecture. Further refining this approach, π_0 [7] incorporates a flow-matching mechanism to generate continuous, precise action trajectories, enhancing fine-grained manipulation skills.

2.3. Video Benchmarks for MLLM

Recently, multiple benchmarks [19, 40, 41, 52] have emerged for comprehensively evaluating MLLMs’ ability in (long) video understanding, especially visual perception and semantic reasoning through Video Question Answering. LongVideoBench [41] and LVBench [40] focus on long video understanding, while Video-MME [19] and MMBench-Video [18] comprehensively evaluate MLLMs across various video-related tasks. Existing benchmarks primarily focus on high-level semantic understanding, such as entity recognition and event understanding, largely confined to a temporal extension of 2D image understanding, lacking precise 3D spatial and temporal reasoning of physical quantities. Recent works such as VSI-Bench [45] have introduced visual-spatial intelligence tasks for MLLMs, where models are required to provide numerical answers in certain scenarios. However, as illustrated in Table 1, the limited inclusion of scenes and spatial-temporal tasks restricts their ability to capture the complexities of the real

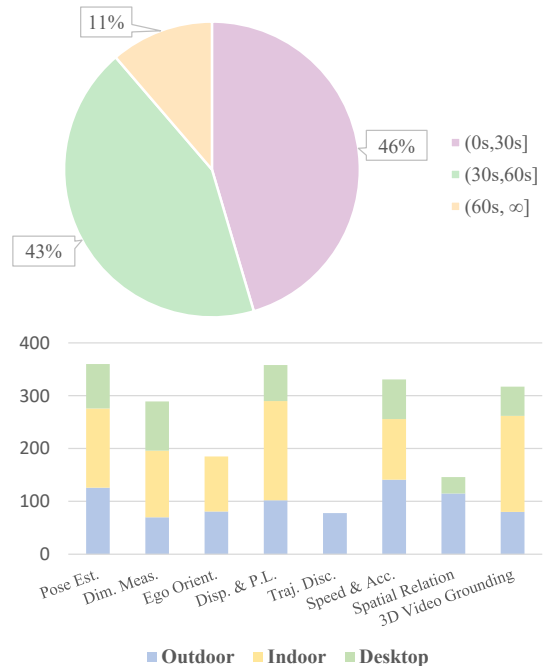


Figure 3. **Benchmark Statistics.** Top: Video length distribution across different categories and datasets. Bottom: The number of questions contributed by each dataset for evaluating different capabilities.

physical world. In contrast, STI-Bench comprehensively evaluates models’ ability in precise spatial-temporal understanding through tasks of static spatial measurement and physical motion understanding in Desktop, Indoor and Outdoor scenarios.

3. STI-Bench

In this section, we present the detailed design and construction of STI-Bench. The construction pipeline is depicted in Figure 4.

3.1. Task Definition

We propose eight tasks examining distinct aspects of MLLMs’ spatial-temporal understanding, divided into Static Understanding and Dynamic Understanding.

Static Understanding

- a. **Dimensional Measurement.** Estimates object geometric size and distances, requiring transformation from 2D pixel observations to physical measurements.
Example: "What is the height of this box?"
- b. **Spatial Relation.** Identifies spatial relationships among objects or between camera and objects, testing relative positioning understanding across viewpoints.
Example: "Is the chair on the left or right of the table?"
- c. **3D Video Grounding.** Retrieves object’s 3D bounding box given semantic descriptions, requiring alignment of linguistic and visual features.
Example: "Locate the 3D bounding box of the red suitcase near the bed."

Dynamic Understanding

- d. **Displacement and Path Length.** Measures travel distance between time points, requiring motion tracking across frames.
Example: "How far has the car traveled from 1s to 18s?"
- e. **Speed and Acceleration.** Computes motion parameters by integrating spatial displacement with time intervals.
Example: "What is the average speed of the camera?"
- f. **Ego-Centric Orientation.** Examines camera’s azimuth orientation changes, requiring understanding of rotation representations.
Example: "How many degrees does the camera’s horizontal orientation shift?"
- g. **Trajectory Description.** Describes motion paths throughout the video, testing ability to abstract spatial motion into language.
Example: "Summarize the camera trajectory, including distances moved and turns made."
- h. **Pose Estimation.** Estimates camera pose at specified time points using RGB data, requiring visual odometry capabilities.
Example: "Given the initial pose, what is the camera’s pose at the requested time?"

These tasks collectively evaluate comprehensive spatial-temporal intelligence across different scales, requiring fundamental 3D spatial reasoning, physical common sense, and cross-modal information integration over time.

3.2. Benchmark Construction

Data Collection. To encompass a broad spectrum of real-world environments, STI-Bench covers three major scenarios: *Desktop*, *Indoor*, and *Outdoor*. Accordingly, we draw from three publicly available datasets—**Waymo** [36] for autonomous driving, **ScanNet**[16] for indoor 3D scene reconstruction, and **Omni6DPose**[51] for desktop-scale 6D object pose estimation. These datasets provide frame-by-frame camera intrinsic and extrinsic parameters, as well as point clouds for each object, which we map to two-dimensional bounding boxes in each frame.

Automatic QA Pair Generation. We used MLLMs to produce detailed semantic descriptions for each object, such as "A beige minivan with a roof rack," "A refrigerator with emoji magnets, photos, and a to-do list," or "A red backpack on a brown leather sofa." Next, leveraging the frame-by-frame annotations, we computed the ground-truth information required for each task. We then provided the ground-truth data, object descriptions, and task-specific QA requirements to MLLMs to generate a diverse set of questions and challenging answer options.

Human Quality Control. During QA pair generation, several issues arose: LLM-generated descriptions could be inaccurate or fail to uniquely identify the target object; some questions and options remained unreasonable or incorrect, even with detailed guidelines; and in certain cases, the video alone did not provide sufficient information (e.g., when the camera was occluded but lidar data were available). To address these challenges, we developed a website for multiple rounds of manual filtering and sampling-based review, ensuring high-quality questions. We also randomly shuffled the answer options to enhance evaluation robustness. Ultimately, we curated more than 2,000 high-quality QA pairs from over 300 videos. Details are shown in Figure 3.

Fine-Grained Adjustment. To ensure task diversity and challenge, each question is equipped with carefully constructed distractor options that are both significantly different from the correct answer while remaining within a reasonable range. Considering the varying precision requirements across different scenarios—millimeter-level for desktop applications, centimeter-to-decimeter-level for indoor environments, and decimeter-to-meter-level for outdoor scenes—we applied scenario-specific scaling factors to generate distractors and employed logarithmic sampling techniques to distribute numerical differences more evenly, thereby enhancing the robustness of our evaluation.

Specifically, given a scene type $S \in \{\text{Desktop}, \text{Indoor}, \text{Outdoor}\}$, a correct answer A_{correct} , and four initial distractor options $\{A_{d1}, A_{d2}, A_{d3}, A_{d4}\}$, our

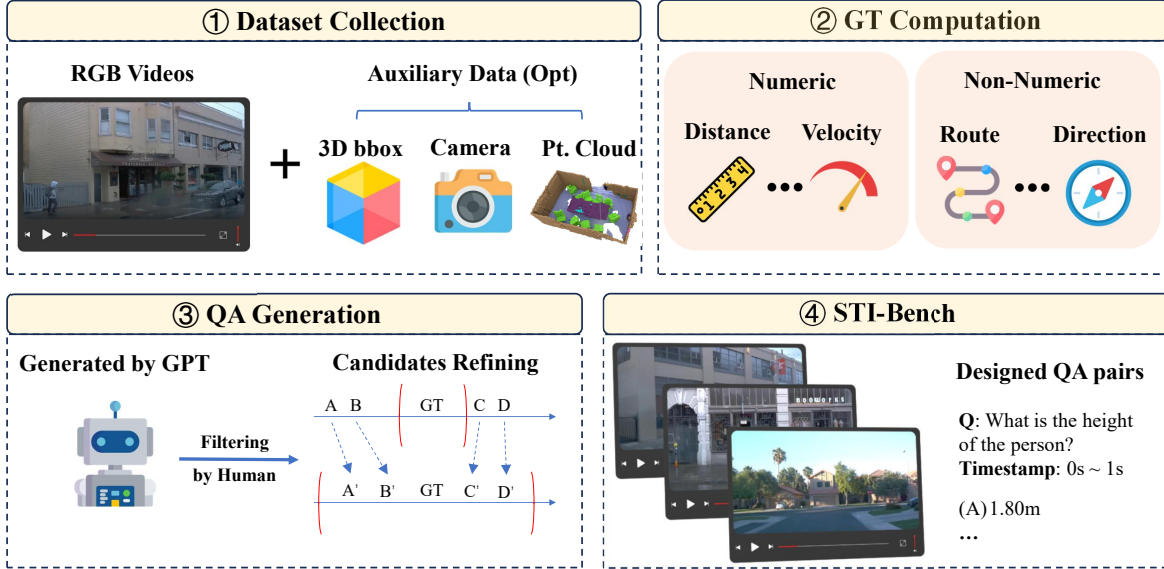


Figure 4. **Benchmark curation pipeline.** The pipeline first aggregates multi-scene RGB datasets that contain 3D bounding box annotations, camera parameters, and point cloud data, which serve as the basis for computing ground truth. From these datasets, we extract numerical ground truth such as distance and velocity, as well as textual descriptions of trajectories and directions. Subsequently, we leverage GPT to assist in generating QA pairs and design a website for rigorous human verification and filtering.

fine-grained adjustment process consists of the following steps:

1. We determine the error range $[E_{min}, E_{max}]$ based on the scene type S : - Desktop: $[0.5 \text{ cm}, 5 \text{ cm}]$ - Indoor: $[5 \text{ cm}, 50 \text{ cm}]$ - Outdoor: $[0.5 \text{ m}, 5 \text{ m}]$

2. We logarithmically sample an error value e from the corresponding error range:

$$e = E_{min} \cdot (E_{max}/E_{min})^u$$

where $u \sim \mathcal{U}(0, 1)$ is a uniform random number.

3. We adjust each distractor option A_{di} to A'_{di} using a weighted average:

$$A'_{di} = (1 - w) \cdot A_{di} + w \cdot A_{correct}$$

where the weight w is identical for all distractors and is determined by solving:

$$\min_{i \in \{1, 2, 3, 4\}} \|A'_{di} - A_{correct}\| = e$$

This yields:

$$w = 1 - \frac{e}{\min_{i \in \{1, 2, 3, 4\}} \|A_{di} - A_{correct}\|}$$

This approach ensures that the minimum distance between any adjusted distractor and the correct answer equals the sampled error value e , while other distractors maintain relatively larger distances but are adjusted using the same weight factor.

4. Experiments

4.1. Settings

We conduct a thorough evaluation of leading MLLMs from diverse model families, focusing on both proprietary and open-source solutions. Specifically, we assess four proprietary models: GPT-4o[33], Gemini-2.0-Flash[38], Gemini-2.5-Pro[38], and Claude-3.7-Sonnet[2], as well as several representative open-source MLLMs that have undergone specialized video-related training, including Qwen2.5-VL-72B[5], InternVL2.5-78B[13], MiniCPM-V-2.6[47], VideoChat-Flash[27], VideoChat-R1[28] and VideoLLaMA3-7B[49].

Considering the stability of open-source models and the API limitations of proprietary models, we uniformly sample 30 frames from the video for each record and explicitly indicate the sampling FPS within the prompt. An exception is made for Claude-3.7-Sonnet, for which only 20 frames are sampled due to its API constraints. Our benchmark tasks are presented in a multiple-choice format with five possible answers, hence a random guess baseline yields 20% accuracy. We measure each model’s accuracy by directly comparing the model’s selected answer with the ground truth.

4.2. Main Results

As shown in Table 2 and Table 3, we present a comprehensive evaluation of various MLLMs on STI-Bench. Gemini-2.5-Pro achieves the highest average accuracy of 41.4%,

Methods	Rank	Avg.	Static Understanding			Dynamic Understanding				
			Dim. Meas.	Spatial Relation	3D Video Grounding	Disp. & P.L.	Speed & Acc.	Ego Orient.	Traj. Desc.	Pose Est.
<i>Proprietary Models (API)</i>										
GPT-4o[33]	8	34.8	27.1	51.8	29.0	23.2	35.4	33.7	32.0	53.6
Gemini-2.0-Flash[38]	4	38.7	31.9	50.0	31.8	27.7	32.1	10.8	38.5	61.3
Claude-3.7-Sonnet[2]	3	40.5	29.8	45.5	35.7	28.9	38.8	40.0	47.4	62.6
Gemini-2.5-Pro[20]	1	41.4	38.7	53.8	36.9	33.9	33.1	52.5	47.4	50.4
<i>Open-source Models</i>										
MiniCPM-V-2.6[47]	10	26.9	27.7	44.5	29.0	19.0	25.7	7.0	30.8	35.6
VideoChat-R1[28]	9	32.8	23.2	47.3	31.5	22.4	31.1	26.0	47.9	48.3
VideoLLaMA3-7B[49]	7	35.2	29.4	48.6	36.1	21.5	36.7	23.2	54.6	48.1
VideoChat-Flash[27]	6	36.3	33.6	51.4	33.1	27.1	32.3	22.2	54.2	51.4
InternVL2.5-78B[13]	5	38.5	29.9	52.8	31.6	24.9	37.2	49.2	43.6	53.6
Qwen2.5-VL-72B[5]	2	40.7	31.5	47.6	39.1	25.1	38.4	43.8	51.3	60.6

Table 2. Evaluation on STI-Bench. Orange = best; Light Orange = second best. Note: The random baseline for all tasks is 20%.

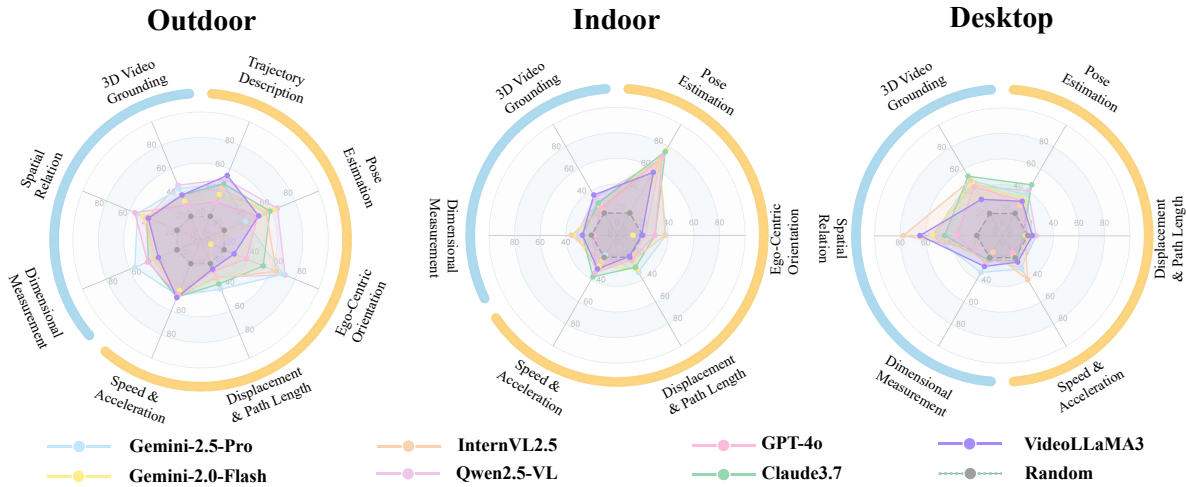


Figure 5. Evaluation results across different scenes and tasks.

Model	Outdoor	Indoor	Desktop	Overall
MiniCPM-V-2.6[47]	27.9	26.1	26.6	26.9
VideoChat-R1.7B[28]	41.1	28.0	27.1	32.8
GPT-4o[33]	41.4	33.1	27.3	35.2
VideoLLaMA3-7B[49]	41.2	31.9	30.5	35.2
Gemini-2.0-Flash[38]	39.1	35.9	30.5	36.0
VideoChat-Flash[27]	38.6	36.7	31.3	36.3
InternVL2.5-78B[13]	46.1	34.5	32.1	38.5
Claude-3.7-Sonnet[2]	47.2	38.2	32.3	40.5
Qwen2.5-VL-72B[5]	50.6	35.0	33.5	40.7
Gemini-2.5-Pro[20]	48.7	37.1	35.8	41.4

Table 3. Evaluation results across different scenes.

closely followed by Qwen2.5-VL-72B at 40.7% among open-source models. While these results significantly ex-

ceed the random guess baseline (20%), they highlight substantial room for improvement in spatial-temporal understanding.

Performance across different scene types shows varied patterns. Qwen2.5-VL-72B demonstrates strong performance in outdoor scenarios (50.6%), while Gemini-2.5-Pro performs well outdoors (48.7%) and leads in desktop scenes (35.8%). In indoor environments, Claude-3.7-Sonnet (38.2%) and Gemini-2.5-Pro (37.1%) lead. These variations suggest different model specializations, possibly influenced by training data distribution.

Task-specific performance reveals particularly challenging areas. Models generally struggle with precise quantitative estimation tasks. The best performance on Dimensional Measurement is achieved by Gemini-2.5-Pro (38.7%), and on Displacement & Path Length also by Gemini-2.5-Pro



Example 1

From 21.0 seconds to 24.0 seconds. What is the average speed of the camera between 21s and 24s?

A.0.13m/s B.0.44m/s C.0.15m/s D.0.42m/s E.0.24m/s

Gemini 2.5 pro:

*Output code estimates the speed of the 'hole_chili' feature between 21s and 24.6s. It uses pre-defined normalized coordinates to calculate displacement. Assuming a board width of 2m for scale, it estimates the real speed (~0.141 m/s). Finally, it finds the multiple-choice option (A-E) with the closest speed value and prints its letter ('C').

Correct!

Example 2

What is the camera's instantaneous speed around t=12s?

A.0.14m/s B.0.09m/s C.0.24m/s D.0.23m/s E.0.27m/s

Gemini 2.5 pro:

*The code estimates angular speed based on visual inspection (~18 deg/s). Assuming a camera orbital radius of 0.75m, it calculates the corresponding linear speed (~0.235 m/s) using speed = angular_speed * radius. It then compares this speed to options A-E and selects the closest one ('D').

Not accurate enough.

Wrong!

Figure 6. The simplified thought process examples of Gemini 2.5 Pro.

(33.9%). In contrast, models demonstrate stronger capabilities in Pose Estimation (best: 62.7% by Claude-3.7-Sonnet) and Spatial Relation tasks (best: 53.8% by Gemini-2.5-Pro).

Among open-source models, Qwen2.5-VL-72B stands out with 40.7% average accuracy, demonstrating highly competitive performance. Other open-source models like InternVL2.5-78B (38.5%) also show promise, though smaller models like MiniCPM-V-2.6 (26.9%) lag behind their larger counterparts.

Even the best-performing model, Gemini-2.5-Pro, achieves only 41.4% average accuracy on our benchmark. These results indicate that current MLLMs, despite impressive general visual understanding capabilities, still require significant advancements in precise spatial-temporal intelligence for embodied tasks.

4.3. Experimental Analysis

Given that Gemini-2.5-Pro exhibits strong multi-modal reasoning capabilities and achieves the top overall performance (41.4% average accuracy), we select it as a representative for in-depth analysis.

Performance varies across scene types: it performs strongest in outdoor scenarios (48.7%), followed by indoor

(37.1%) and desktop environments (35.8%). This disparity might suggest that the model's training data is better attuned to outdoor and larger-scale understanding.

Examining task-specific performance, we observe its strongest capabilities in tasks involving relative understanding and object state. It achieves high scores in Spatial Relation (53.8%) and Ego-Centric Orientation (52.5%), followed by Pose Estimation (50.4%). While tasks requiring precise quantitative estimation show lower absolute scores, Gemini-2.5-Pro performs competitively and often leads in these areas, achieving top scores in Dimensional Measurement (38.7%) and Displacement & Path Length (33.9%).

By leveraging the model's reasoning process and uniformly sampling approximately 200 error records across each task type and scenario, we categorize errors into three representative patterns. Figure 7 shows the distribution of error categories.

Inaccurate Spatial Quantification The model faces significant challenges in accurately estimating static spatial properties and relationships from visual inputs. These difficulties manifest when estimating object dimensions, distances between objects or between camera and objects, and absolute 3D coordinates at specific time points. These er-

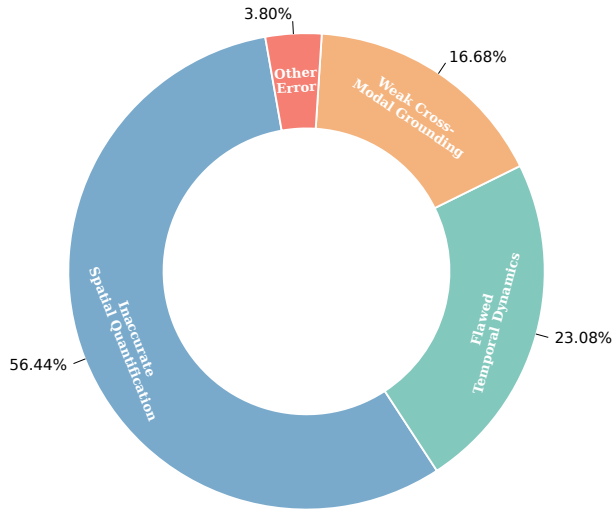


Figure 7. Distribution of error categories in Gemini-2.5-Pro across our sampled error cases.

rors stem from a lack of clear visual size references, difficulty distinguishing between numerically close options, and the inherent challenges of inferring metric scale from 2D pixels and estimating depth with monocular cameras. Such limitations directly impact performance in dimensional measurement, spatial relation, and 3D video grounding tasks.

Flawed Temporal Dynamics Understanding The model struggles to perceive, track, and interpret cross-frame information that changes over time, such as motion and its dynamics. This results in erroneous calculations or descriptions of displacement, path length, speed, acceleration, directional changes, and overall trajectory shapes. The model particularly struggles with relative motion (distinguishing object motion from camera motion), a problem exacerbated by sparse temporal sampling. These difficulties arise from challenges in integrating information across frames, lack of internal models for physics/kinematics, inability to separate ego-motion from object motion, and information loss due to sparse sampling. These issues manifest in tasks involving displacement and path length, speed and acceleration, ego-centric orientation, trajectory description, and pose estimation.

Weak Cross-Modal Grounding and Integration The model fails to properly connect textual queries/instructions with relevant spatial-temporal visual content, or to integrate provided non-visual data (such as initial poses) with visual information. This includes misinterpreting temporal constraints (like "from 1s to 18s," "at the end," "the mo-

ment of last co-occurrence"), failing to correctly utilize provided initial conditions (e.g., initial camera pose in pose estimation tasks), and incorrectly associating structured data (coordinates, timestamps) with visual elements. These errors stem from deficiencies in parsing structured/natural language instructions and difficulty integrating information from different modalities into a unified reasoning process. This affects all tasks that rely on specific instructions or initial data.

These error patterns highlight that, despite Gemini-2.5-Pro's strong performance relative to other models, it still faces significant challenges in precise spatial-temporal understanding. Its limitations in quantitative estimation and complex spatial-temporal reasoning indicate that current MLLMs remain far from achieving the reliability required for embodied AI or autonomous driving applications.

5. Conclusion

We introduced STI-Bench, a comprehensive benchmark to assess MLLMs' spatial-temporal understanding through over 300 real-world videos and 2,000 QA pairs of robot outdoor, indoor, and desktop scenarios, which reveals significant limitations in current MLLMs' spatial-temporal understanding capabilities, with top-performing models like Gemini-2.5-Pro achieving around 41.4% average accuracy. Models particularly struggle with precise quantitative tasks like dimensional measurement. Our analysis identifies three key weaknesses: inaccurate spatial quantification, flawed temporal dynamics understanding, and weak cross-modal integration. These findings emphasize the substantial gap between current capabilities and the reliability needed for embodied AI and autonomous driving applications. STI-Bench provides a valuable framework for evaluating and improving MLLMs' ability to understand the physical world—essential for developing the next generation of embodied intelligent systems.

Acknowledgement

This work is funded by NSFC-62306046. We thank the data annotation team of MolarData for their support, as well as the volunteer contributors Yuxin Liu, Junjie Ruan, Xucheng Liao, Zixuan Huang, Tianrui Wan, Qingcheng Wei, Yujie Yao, Shangyang Dong, Zhuofan Zeng and Yiming Li for their valuable work as human annotators and evaluators.

References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1, 2

- [2] Anthropic. Claude 3.7 Sonnet and Claude Code. Anthropic News Announcement, 2025. Announcement on Anthropic blog, Feb 24, 2025. [5](#), [6](#)
- [3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. [1](#)
- [4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. [1](#), [2](#)
- [5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao 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, 2025. [5](#), [6](#)
- [6] Youssef Benchechroun, Megi Dervishi, Mark Ibrahim, Jean-Baptiste Gaya, Xavier Martinet, Grégoire Mialon, Thomas Scialom, Emmanuel Dupoux, Dieuwke Hupkes, and Pascal Vincent. Worldsense: A synthetic benchmark for grounded reasoning in large language models. *arXiv preprint arXiv:2311.15930*, 2023. [3](#)
- [7] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0 : A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. [1](#), [3](#)
- [8] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. [3](#)
- [9] Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*, 2024. [1](#), [2](#)
- [10] Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, et al. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024. [3](#)
- [11] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 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, 2024. [2](#)
- [12] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. [1](#), [2](#)
- [13] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. [5](#), [6](#)
- [14] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision language model. *arXiv preprint arXiv:2406.01584*, 2024. [2](#)
- [15] Zhili Cheng, Yuge Tu, Ran Li, Shiqi Dai, Jinyi Hu, Shengding Hu, Jiahao Li, Yang Shi, Tianyu Yu, Weize Chen, et al. Embodiedeval: Evaluate multimodal llms as embodied agents. *arXiv preprint arXiv:2501.11858*, 2025. [1](#), [3](#)
- [16] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. [2](#), [4](#)
- [17] Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. *arXiv preprint arXiv:2406.05756*, 2024. [3](#)
- [18] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *arXiv preprint arXiv:2406.14515*, 2024. [1](#), [3](#)
- [19] 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. [1](#), [3](#)
- [20] Google DeepMind. Gemini 2.5 pro preview model card, 2025. [6](#)
- [21] 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. [1](#), [3](#)
- [22] Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, Yin Zhou, James Guo, Dragomir Anguelov, and Mingxing Tan. Emma: End-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2410.23262*, 2024. [1](#)
- [23] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabriela Levine, Wensi Ai, Benjamin Martinez, et al. Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation. *arXiv preprint arXiv:2403.09227*, 2024. [1](#)
- [24] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao.

- Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 2
- [25] 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. 1
- [26] Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Li Erran Li, Ruohan Zhang, et al. Embodied agent interface: Benchmarking llms for embodied decision making. In *NeurIPS 2024*, 2024. 1, 3
- [27] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhao Zhu, Haiyan Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024. 5, 6
- [28] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025. 5, 6
- [29] Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 1
- [30] Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection, 2024. 2
- [31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 2
- [32] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters (RA-L)*, 7(3):7327–7334, 2022. 1
- [33] OpenAI. GPT-4o system card, 2024. 1, 5, 6
- [34] Arijit Ray, Jiafei Duan, Reuben Tan, Dina Bashkurova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, Kuo-Hao Zeng, et al. Sat: Spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024. 3
- [35] Yan Shu, Peitian Zhang, Zheng Liu, Minghao Qin, Junjie Zhou, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. *arXiv preprint arXiv:2409.14485*, 2024. 1
- [36] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Etinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 4
- [37] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 2
- [38] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1, 2, 5, 6
- [39] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1, 2
- [40] Weihao 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. 1, 3
- [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, 2025. 1, 3
- [42] Yi Xu, Yuxin Hu, Zaiwei Zhang, Gregory P. Meyer, Siva Karthik Mustikovela, Siddhartha Srinivasa, Eric M. Wolff, and Xin Huang. Vlm-ad: End-to-end autonomous driving through vision-language model supervision. *arXiv preprint arXiv:2412.14446*, 2024. 1
- [43] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 2024. 1
- [44] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 1, 2
- [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. 3
- [46] Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*, 2025. 1, 3
- [47] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 5, 6
- [48] Jianhao Yuan, Shuyang Sun, Daniel Omeiza, Bo Zhao, Paul Newman, Lars Kunze, and Matthew Gadd. Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model. *arXiv preprint arXiv:2402.10828*, 2024. 1
- [49] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang,

- Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multi-modal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. [5](#), [6](#)
- [50] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multi-modal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. [1](#)
- [51] Jiayao Zhang, Weiyao Huang, Bo Peng, Mingdong Wu, Fei Hu, Zijian Chen, Bo Zhao, and Hao Dong. Omni6dpose: A benchmark and model for universal 6d object pose estimation and tracking. In *European Conference on Computer Vision*, pages 199–216. Springer, 2024. [2](#), [4](#)
- [52] 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. [1](#), [3](#)