MMSI-VIDEO-BENCH: A HOLISTIC BENCHMARK FOR VIDEO-BASED SPATIAL INTELLIGENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial understanding over continuous visual input is crucial for MLLMs to evolve into general-purpose assistants in physical environments. Yet there is still no comprehensive benchmark that holistically assesses the progress toward this goal. In this work, we introduce **MMSI-Video-Bench**, a fully human-annotated benchmark for video-based spatial intelligence in MLLMs. It operationalizes a four-level framework, Perception, Planning, Prediction, and Cross-Video Reasoning, through 1,106 questions grounded in 1,278 clips from 25 datasets and in-house videos. Each item is carefully designed and reviewed by 3DV experts with explanatory rationales to ensure precise, unambiguous grounding. We evaluate 25 strong open-source and proprietary MLLMs, revealing a striking human–AI gap: many models perform near chance, and the best reasoning model lags humans by nearly 60%. We further find that spatially fine-tuned models still fail to generalize effectively on our benchmark. Fine-grained error analysis exposes systematic failures in geometric reasoning, motion grounding, long-horizon prediction, and cross-video correspondence. We also show that typical frame-sampling strategies transfer poorly to our reasoning-intensive benchmark, and that neither 3D spatial cues nor chain-of-thought prompting yields meaningful gains. We expect our benchmark to establish a solid testbed for advancing video-based spatial intelligence.

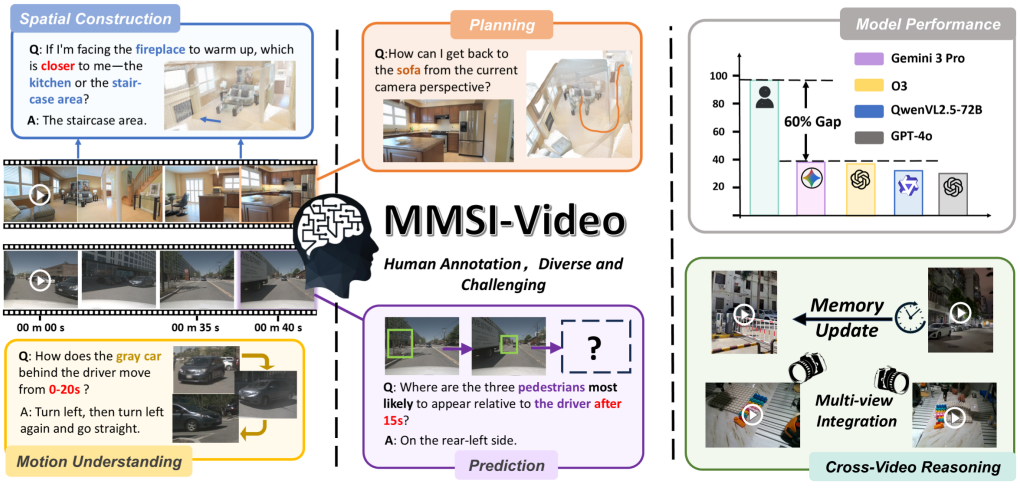

Figure 1: **MMSI-Video-Bench** is a diverse, human-annotated, and challenging benchmark, designed to evaluate models' video-based spatial intelligence, including their ability to perceive, understand, reason, and make decisions over spatio-temporal information in videos. The bar chart in the top-right corner illustrates the substantial performance gap between state-of-the-art models and human performance.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison of MMSI-Video-Bench with other spatial reasoning benchmarks, highlighting its diversity, comprehensiveness, and challenge. INST. and CAM. refer to instance and camera, SU., MU., and CV. refer to Spatial Understanding, Motion Understanding, and Cross-Video Reasoning, respectively. Sources denotes the number of data sources covered, and the Human–AI Gap indicates the performance difference as a percentage.

Benchmark	Sources	Modality	Annotation	Task	Samples Num	Human-AI Gap
SpatialRGPT (Cheng et al., 2024)	-	Single-Image	Auto	local-SU.	1,406	<42
SpatialVLM (Chen et al., 2024)	-	Single-Image	Auto&Human	local-SU.	546	<30
CVBench (Tong et al., 2024)	3	Single-Image	Auto	local-SU.	2,638	-
MultiSPA (Xu et al., 2025)	3	Multi-Image	Auto	local-SU. & short-MU.	7800	-
All-Angles-Bench (Yeh et al., 2025)	2	Multi-Image	Human	SU.	2100	21.2
MMSI-Bench (Yang et al., 2025)	8	Multi-Image	Human	local-SU. & short-MU.	1,000	55.3
VSI-Bench (Yang et al., 2024)	3	Video	Auto	SU. & Plan.	5,000	33
OST-Bench (Lin et al., 2025)	3	Video	Auto	SU. & CAM.-MU.	10,000	29.3
SPAR-Bench (Zhang et al., 2025a)	3	Video	Auto	SU. & CAM.-MU.	7207	27.8
STI-Bench (Li et al., 2025b)	3	Video	Auto	SU. & INST./CAM.-MU.	2,064	-
EgoExoBench (He et al., 2025)	6	Multi-Video	Auto&Human	CV.	7,330	41.9
MMSI-Video (ours)	26	Video/Multi-Video	Human	SU. & MU. & Plan./Pred. & CV.	1,106	58.4

1 INTRODUCTION

For decades, humans have aspired to build an embodied general-purpose AI assistant, akin to JARVIS in *Iron Man*. As MLLMs (Bai et al., 2025; Zhu et al., 2025; Zhang et al., 2025b; Xu et al., 2025; Chen et al., 2025) have begun to exhibit strong language and visual intelligence, they are increasingly viewed as a promising foundation for embodied AGI. One of the most important remaining challenges in this pursuit is to endow MLLMs with spatial intelligence-the ability to perceive, reason about, and interact with physical space from continuous visual inputs, as humans do.

To reliably measure progress on this research, we need rigorous benchmarks. However, existing benchmarks have significant limitations: most operate on discrete single images (Kirillova et al., 2022; Chen et al., 2024; Cheng et al., 2024) or multiple images (Fu et al., 2024b; Yang et al., 2025; Jia et al., 2025) rather than videos, leaving an input gap to the practical setting. Recent video-based benchmarks (Yang et al., 2024; Zhang et al., 2025a; Li et al., 2025b; Lin et al., 2025) also suffer from the following issues: (1) question types are not sufficiently holistic; (2) they rely heavily on templated automatic question generation, which restricts question diversity and may introduce template overfitting or biases (Yang et al., 2025); and (3) data sources and scenes are not comprehensive enough.

In this work, we introduce **MMSI-Video-Bench** to fill these gaps. We build our benchmark around a holistic, multi-level framework for video-based spatial intelligence consisting of Perception, Planning, Prediction, and Cross-Video Reasoning (see 1). Models are required to reason over a single video for spatial perception, capturing global scene information (Spatial Construction) and ego-/exo Motion dynamics. Beyond perception, models should be able to make decisions or take actions to interact with the environment (Planning) and further make Predictions about future spatial states of the world. Finally, a general spatial intelligence model should be capable of Cross-Video Reasoning for multi-view integration and memory updating.

We instantiate the above holistic framework as a diverse, accurate, and challenging MCQ benchmark. We adopt a fully human-designed protocol following (Yang et al., 2025). Eleven 3DV researchers manually design each sample, including selecting a video clip from a curated pool of about 20K videos, designing a novel question, writing the correct answer, distractors, and a brief rationale. A multi-stage review process leverages these rationales to ensure accuracy and unambiguity. Our video pool combines 25 open-source datasets with newly recorded in-house videos, covering a wide range of scenarios, including indoor scans, outdoor driving, robotics, etc. In total, with 400+ hours of annotation and verification, we obtain 1,106 questions grounded in 1,278 video clips, grouped into five task categories and 13 subtypes. Thanks to the diversity of data sources and the holistic coverage of task types in MMSI-Video-Bench, we are also able to build three domain-oriented sub-benchmarks: Indoor Scene Perception Bench, Robot Bench, and Grounding Bench, enabling targeted assessment of specific model capabilities.

Using MMSI-Video-Bench, we conduct a comprehensive evaluation of open-source and proprietary MLLMs. Current models remain far from the desired level of video-based spatial intelligence: many perform close to random guessing, and the best model, Google's Gemini 3 Pro, still trails humans by nearly 60%. To the best of our knowledge, our benchmark yields the largest human–AI performance gap among existing video-based spatial benchmarks. In addition, we also evaluate models that have been spatially fine-tuned, yet their capabilities still fail to generalize effectively on our benchmark, further underscoring its challenge (Sec. 4.3).

To provide diagnostic signals for future research, we perform per-category error analysis (Sec. 5). For Spatial Construction, failures are dominated by geometric reasoning errors; for Motion, by fine-grained grounding failures on long-duration motion; for Planning and Prediction, by prompt–evidence misalignment where models ignore video cues; and for Cross-Video Reasoning, by difficult grounding and matching correspondences across videos. Beyond analysis, our preliminary exploration shows that neither 3D spatial cues nor chain-of-thought prompting yields meaningful gains on MMSI-Video-Bench (Appendix D). We further observe that, due to the reasoning-intensive nature of our benchmark, the commonly used frame-sampling strategy AKS (Tang et al., 2025b) does not transfer directly to MMSI-Video-Bench, leading to additional performance degradation (Appendix C.2). Overall, current MLLMs still have substantial room for improvement in spatial intelligence, and our bench offers an accurate and challenging testbed with actionable insights for advancing video-based spatial reasoning.

2 Related Work

Video-based Benchmarks. Video-based benchmarks evaluate a model's ability to perceive, understand, and reason over temporal and visual information. Early video benchmarks such as MSVD-QA, MSRVTT-QA (Xu et al., 2017), and ActivityNet-QA (Yu et al., 2019) mainly assess global visual comprehension, with limited attention to temporal understanding. Later works like NeXT-QA (Xiao et al., 2021) and MVBench (Li et al., 2024) emphasize temporal dynamics. Subsequently, benchmarks such as LongVideoBench (Wu et al., 2024) and Video-MME (Fu et al., 2024a) further extended evaluation beyond surface-level perception, incorporating temporal–event reasoning. With the advancement of MLLMs, more video-based reasoning benchmarks have emerged, each designed to probe specific aspects of real-world understanding. These include benchmarks focusing on such as complex spatial reasoning within videos (Li et al., 2025b; Yang et al., 2024; Zhang et al., 2025a; Lin et al., 2025), online inference under continuous observation (Lin et al., 2025; Li et al., 2025a), and cross-video reasoning (He et al., 2025). Different from previous benchmarks, our proposed MMSI-Video-Bench serves as a more holistic and challenging benchmark for video spatial intelligence, covering complex reasoning about spatial layouts, motion understanding, and decision-making, as well as reasoning across multiple videos.

Spatial Intelligence Benchmarks. Existing benchmarks for evaluating spatial intelligence in multimodal large language models (MLLMs) vary substantially in task design, modality, scene scope, and the specific spatial abilities they aim to assess. Early benchmarks such as SpatialRGPT (Cheng et al., 2024), SpatialVLM (Chen et al., 2024), and CVBench (Tong et al., 2024) focus on single-image spatial reasoning, emphasizing depth and distance. Later, video-based benchmarks, VSI-Bench (Yang et al., 2024), SPAR-Bench (Zhang et al., 2025a), and OST-Bench (Lin et al., 2025), extend this to indoor scenes with object–object and object–camera relations, yet remain constrained by limited scene diversity and static spatial contexts. More recent benchmarks, including MultiSPA (Xu et al., 2025), MMSI-Bench (Yang et al., 2025), and STI-Bench (Li et al., 2025b), incorporate dynamic environments and cover a wider range of indoor–outdoor scenarios. MultiSPA and MMSI-Bench serve as demanding benchmarks focusing on multi-image local spatial localization and short-term motion understanding, whereas STI-Bench targets numerical estimation of instance-centric spatial states and motion trajectories within videos. Nevertheless, existing benchmarks focus on limited aspects or scene types, lacking a holistic evaluation across diverse real-world contexts. In contrast, our MMSI-Video-Bench, curated from diverse real-world videos and fully human-annotated, offers a more holistic and realistic assessment of spatial intelligence in MLLMs.



Figure 2: Illustrative examples of different subtypes in MMSI-Video-Bench. Rel., Inst., and Cam. stand for Relationship, Instance, and Camera.

3 MMSI-VIDEO-BENCH

As a video-based spatial intelligence benchmark, MMSI-Video-Bench primarily evaluates a model's capability to perceive, understand, and reason over video information, encompassing two core dimensions: **Spatial.** This dimension concerns the spatial states (e.g., position, shape) of entities such as instances, scenes, or cameras, and their spatial relations (e.g., front–back, left–right, near–far) at a fixed moment. To assess this ability, we introduce the *Spatial Construction* category, which requires the model to infer fine-grained global spatial layouts from partial and sequential video observations.

Spatio-Temporal. When motion occurs in the video (from the camera or instances), the spatial configuration changes over time. The *Motion Understanding* category evaluates the model's ability to reason about long-term motion dynamics across consecutive frames, including camera motion, individual instance motion, and interactive motion arising from interactions among multiple instances.

After understanding spatio-temporal information, the next step is decision-making based on video understanding. *Planning* and *Prediction* categories focus on higher-level reasoning over sparse video information: Planning tasks require the model to devise actions toward a specific goal based on

visual cues; Prediction tasks test the model's ability to infer or imagine outcomes under hypothetical conditions.

The above categories comprehensively cover spatial intelligence within a single video. However, real-world understanding requires reasoning across multiple videos. To achieve a more holistic evaluation, we extend MMSI-Video-Bench with *Cross-Video Reasoning*, including: (1) Memory Update. From a temporal perspective, observations of the same scene in real-world settings are often temporally discontinuous (i.e., we do not continuously stay in one place). The model must therefore retain contextual memory from past observations and update it as new information becomes available. (2) Multi-View Integration. From a spatial perspective, a single viewpoint rarely captures the complete spatial information of a complex scene. The model must thus integrate observations from multiple viewpoints to construct a unified representation of the scene. Fig. 2 illustrates example cases for each subtype. In A.1, we present a summary table that outlines the details of each subtype.

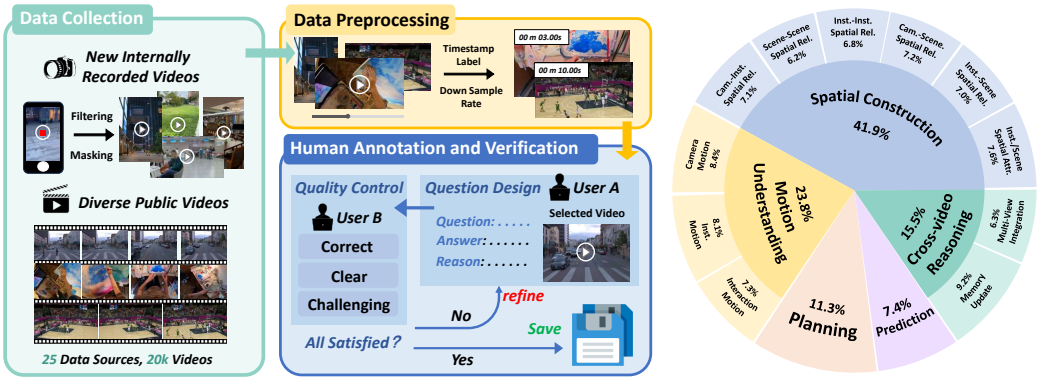

Figure 3: (Left) The pipeline of benchmark sample construction. (Right) The distribution of question types in MMSI-Video-Bench.

3.1 Benchmark Construction

Data Collection & Preprocessing. To ensure diversity in our benchmark, we curated videos spanning a broad spectrum of real-world scenarios. The collection includes a wide range of capture types, such as tabletop recordings, indoor scenes from single-room to multi-floor environments, outdoor building and street views, natural landscapes, sports activities, and movie footage. Our data sources include 25 publicly available datasets (listed in Appendix A.2), amounting to approximately 20k video clips. In addition, we manually recorded and carefully selected 140 supplemental in-house videos, each anonymized via masking to protect personal privacy. All videos were then downsampled to an appropriate frame rate for each category such that no key information would be lost. Each frame was timestamped in the top-left corner ("xx min yy s") to facilitate precise temporal referencing during question design.

Human Annotation. As shown in Fig. 3, all annotations were conducted by a team of eleven trained researchers specializing in 3D vision. To preserve data diversity, annotation tasks were assigned so that each annotator received a balanced mix of task categories. We developed a dedicated annotation interface that allowed annotators to view all videos and design questions directly within the interface. Based on the provided visual context, annotators composed questions, corresponding answers, and concise reasoning, particularly for challenging items, to facilitate subsequent verification. All questions were designed as multiple-choice items, containing between four and six answer options and optionally adopting an interleaved image–text format.

Data Quality Control. We implemented a strict acceptance protocol, where eleven researchers performed cross-evaluation, reviewing each other's work. Each annotation had to meet three criteria: clear (the question is unambiguous and clearly stated), correctness (the answer is unique and factually accurate), and challenge (the question requires non-trivial reasoning). Only annotations that met all three criteria were accepted, with a 100% approval rate required. Annotations that failed were revised based on feedback and resubmitted for reevaluation.

Table 2: Performance of various models under the Sufficient-Coverage and Uniform-50 settings. The highest and second-highest average scores across all settings are highlighted in dark green and light green, respectively. Attr. denotes Attribute; Inst. (I) denotes Instance; Cam. (C) denotes Camera; Scen. (S) denotes Scene; and Inter. denotes Interaction. MU. and MV. represent Memory Update and Multi-View Integration, respectively.

Models	Spatial Construction						Motion Understanding			Cross-Video		Plan.	Pred.	Avg.
	Attr.	I-I	I–S	S–S	C-I	C–S	Cam.	Inst.	Inter.	MU.	MV.	-	-	
(Sufficient-Coverage) Proprietary														
O4-mini	31.3	35.5	39.0	30.4	36.7	36.2	**45.2**	28.9	30.9	40.2	31.4	**38.7**	26.8	35.1
O3	39.8	31.6	**46.8**	**36.2**	**43.0**	36.2	34.4	**34.4**	38.3	**44.1**	34.3	34.7	**31.7**	**37.3**
GPT-4o	22.9	30.3	35.1	26.1	26.6	27.5	24.7	28.9	24.7	34.3	25.7	29.8	26.8	28.1
Gemini 2.5 Flash	**47.0**	36.8	40.3	31.9	30.4	**48.8**	38.7	30.0	**43.2**	37.2	30.0	32.3	30.5	36.6
-Thinking	43.4	**38.2**	41.6	24.6	34.2	36.2	44.1	**34.4**	35.8	36.3	**38.6**	36.3	**31.7**	36.7
(Sufficient-Coverage) Open-source														
InternVL2.5-8B	26.5	27.6	26.0	**36.2**	19.0	30.0	29.0	33.3	30.9	34.3	22.9	24.2	32.9	28.7
InternVL3-8B	28.9	**38.2**	27.3	31.9	22.8	23.8	36.6	31.1	30.9	30.4	34.3	26.6	23.2	29.6
InternVideo2.5-8B	26.5	26.3	23.4	27.5	22.8	28.8	26.9	23.3	25.9	34.3	21.4	29.0	29.3	26.9
QwenVL2.5-7B	31.3	21.1	19.5	30.4	19.0	31.2	35.5	**36.7**	29.6	27.4	30.0	26.6	**35.4**	28.8
QwenVL2.5-32B	**33.7**	28.9	31.2	31.9	**30.4**	30.0	**43.0**	35.6	25.9	25.5	**41.4**	33.1	30.5	**32.4**
QwenVL2.5-72B	**33.7**	22.4	28.6	30.4	24.1	28.8	32.3	33.3	39.5	**35.3**	35.7	**35.5**	30.5	31.8
QwenVL3-8B	27.8	31.5	**33.3**	26.9	23.1	22.4	27.3	29.6	**41.0**	28.4	29.9	26.5	32.0	29.1
QwenVL3-30B	31.3	28.9	26.0	24.6	22.8	**36.2**	29.0	21.1	33.3	29.4	28.6	30.6	**35.4**	29.1
-Thinking	27.7	36.8	29.9	33.3	**30.4**	25.0	26.9	23.3	30.9	25.5	25.7	26.6	25.6	28.0
(Uniform-50) Proprietary														
Claude-haiku-4.5	31.3	31.6	32.5	34.8	32.9	35.0	35.5	**41.1**	35.8	33.3	**38.6**	31.4	32.9	34.3
GPT-5	44.6	38.2	**44.2**	**39.1**	40.5	41.2	37.6	28.9	**42.0**	33.3	28.6	34.7	28.1	36.8
O4-mini	36.1	34.2	36.4	33.3	39.2	42.5	37.6	32.2	37.0	32.4	30.0	28.2	28.1	34.2
O3	34.9	34.2	40.3	36.2	40.5	43.8	37.6	37.8	35.8	**41.2**	25.7	**41.1**	26.8	37.0
GPT-4o	26.5	26.3	31.2	26.1	38.0	38.8	36.6	35.6	27.2	30.4	28.6	33.1	29.3	31.6
Gemini 3 Pro	44.6	**39.5**	**44.2**	33.3	29.1	43.8	35.5	40.0	38.3	34.3	37.1	38.7	**35.4**	**38.0**
Gemini 2.5 Flash	37.4	38.2	40.3	30.4	**43.0**	43.8	39.8	31.1	38.3	30.4	37.1	30.6	24.4	35.4
-Thinking	43.4	25.0	42.9	23.2	32.9	**47.5**	38.7	28.9	38.3	31.4	28.6	40.3	31.7	35.2
Doubao-1.5-thinking	**45.1**	26.7	18.8	38.6	29.3	32.7	33.3	32.2	40.0	31.4	30.8	32.0	20.4	31.6
Seed-1.6-vision	43.1	26.7	16.7	38.6	34.5	40.0	**40.9**	35.6	40.0	32.9	38.5	37.3	25.9	34.9
(Uniform-50) Open-source														
InternVL2.5-8B	21.7	36.8	31.2	30.4	17.7	31.2	32.3	27.8	30.9	35.3	28.6	26.6	28.1	29.1
InternVL2.5-38B	27.7	30.3	26.0	27.5	11.4	33.8	35.5	34.4	34.6	35.3	28.6	33.1	**40.2**	31.0
InternVL2.5-78B	33.7	36.8	**37.7**	18.8	29.1	32.5	28.0	34.4	30.9	31.4	28.6	**35.5**	26.8	31.4
InternVL3-8B	33.7	32.9	29.9	27.5	19.0	18.8	**36.6**	32.2	37.0	31.4	32.9	33.1	26.8	30.4
InternVL3-38B	33.7	36.8	31.2	24.6	16.5	36.2	30.1	30.0	27.2	30.4	27.1	24.2	28.1	28.8
InternVL3-78B	**43.4**	31.6	35.1	23.2	**30.4**	31.2	34.4	32.2	38.3	37.2	34.3	27.4	24.4	32.5
InternVideo2.5-8B	27.7	32.9	22.1	23.2	22.8	28.8	25.8	30.0	25.9	27.4	34.3	29.8	24.4	27.4
LLaVA-Video-7B	27.7	28.9	28.6	30.4	19.0	25.0	31.2	35.6	30.9	36.3	20.0	21.0	35.4	28.5
LLaVA-Video-72B	36.1	**39.5**	28.6	21.7	15.2	23.8	30.1	**42.2**	37.0	28.4	32.9	29.0	29.3	30.4
QwenVL2.5-7B	26.5	25.0	29.9	**34.8**	20.2	**37.5**	33.3	31.1	34.6	24.5	22.9	34.7	28.1	29.7
QwenVL2.5-32B	25.3	26.3	31.2	20.3	21.5	30.0	33.3	35.6	25.9	28.4	32.9	27.4	31.7	28.6
QwenVL2.5-72B	32.5	18.4	**37.7**	29.0	29.1	28.8	34.4	41.1	**40.7**	36.3	**41.4**	26.6	30.5	**32.7**
QwenVL3-8B	33.7	25.0	24.7	26.1	24.1	30.0	35.5	22.2	29.6	31.4	20.0	25.0	29.3	27.6
QwenVL3-30B	27.7	31.6	33.8	20.3	19.0	26.2	29.0	28.9	32.1	31.4	31.4	29.0	31.7	28.8
-Thinking	27.7	36.8	**37.7**	31.9	25.3	31.2	26.9	26.7	28.4	**39.2**	30.0	28.2	31.7	30.8
Baseline														
Random Guessing	24.1	23.7	24.4	24.4	24.2	24.3	24.1	24.9	24.8	23.1	23.2	24.8	24.4	24.1
Human Level	95.2	94.8	96.3	96.0	92.8	94.9	96.8	95.6	93.7	94.4	92.0	95.1	94.2	96.4

Statistics. After the construction of benchmark samples, the final MMSI-Video-Bench contains 1,106 questions grounded in 1,278 video clips, covering five main categories and 13 subtypes, with their distribution shown in Fig. 3. The average video duration is 1 minute 12 seconds, and the average question length is 164.5 characters. Further statistical results are presented in Appendix A.4.

4 Experiments

4.1 Evaluation Settings

We evaluate a wide range of state-of-the-art open-source and proprietary models on MMSI-Video-Bench, including GPT-5/ O3/ O4-mini/ GPT-4o, Gemini 3 Pro/ Gemini 2.5 Flash, Claude-4.5, Seed-1.6-Vision, Doubao-1.5-thinking, InternVL series, QwenVL series, LLaVA-Video series, and others.

All proprietary models are tested via their official APIs, while all open-source models are evaluated using 8×A100 GPUs and follow their officially released inference configurations. Due to (i) the limited number of images allowed per request by some proprietary APIs, and (ii) out-of-memory issues when running large open-source models on long videos, we establish two evaluation tracks:

Uniform-50. Each model receives exactly 50 uniformly sampled frames from the original video. This configuration aligns with the recommended number of input frames for most evaluated models.

Sufficient-Coverage. Each model is provided with the complete set of image frames used during the question design process, ensuring that no visual information is omitted.

For comparison, we additionally provide two baselines: random guessing and human performance. Since all questions in MMSI-Video-Bench are multiple-choice, we adopt an exact-match accuracy metric, where a prediction is considered correct only if it exactly matches the ground-truth.

4.2 Main Results

We report the performance of various models on our benchmark. Tab. 2 summarizes the results of all evaluated models across different question subtypes under two evaluation settings. From the results in the table, we can draw the following key observations:

Substantial gap between MLLMs' and humans' performance. Model performance across all question types in MMSI-Video-Bench falls significantly short of human-level performance. Most models achieve low scores, some approaching the level of random guessing, and even the best-performing model, Gemini 3 Pro (38.0), lags behind humans (96.4) by nearly 60%. This indicates current models are unable to handle these challenging tasks, highlighting the inherent difficulty of MMSI-Video-Bench. This observation motivates a deeper investigation into the underlying reasons for the models' subpar performance.

MLLMs exhibit weaknesses across all categories. Previous spatial reasoning benchmarks have primarily focused on evaluating models' Spatial Construction abilities across various scenarios and contexts, consistently showing poor performance in this aspect (Yang et al., 2024; 2025; Lin et al., 2025). Our benchmark provides a more holistic evaluation, extending beyond Spatial Construction to assess other dimensions such as Motion Understanding, Planning, Prediction, and Cross-Video Reasoning. As shown in Tab. 2, we observe that models also struggle considerably in these areas. In the subsequent error analysis section, we further investigate the specific capability bottlenecks underlying these weaknesses.

Sufficient-Coverage does not outperform Uniform-50. The Sufficient-Coverage setting is assumed to result in better performance than Uniform-50, as it provides the most visual information. However, we observe that most models show no significant improvement, and in many cases, even a performance drop under Sufficient-Coverage. Prior work (Wang et al., 2025b) has likewise shown that more input frames can introduce redundancy that hinders reasoning. To further enhance model performance, it is crucial to develop more effective strategies for key frame sampling. An in-depth analysis of the impact of sampling strategies on model performance is provided in Appendix C.2.

Comparison across models. We observe that proprietary models consistently outperform open-source ones. Among open-source models, the best-performing ones, QwenVL2.5-72B (Uniform-50, 32.7) and QwenVL2.5-32B (Sufficient-Coverage, 32.4), still exhibit a noticeable gap compared to most proprietary models under the same settings. Within open-source models, the results broadly follow the trend that larger parameter scales lead to better performance. In contrast, enabling the thinking mode brings only marginal improvements, as shown by comparing Gemini 2.5 Flash and QwenVL3-30B with their thinking mode versions.

Table 3: Performance of various spatially fine-tuned models and their corresponding base models.

	LLaVA-Video-7B	Spatial-MLLM	VLM3R	QwenVL2.5-3B	SpaceQwen
Modification	-	Architecture	Architecture	-	Fine-tuning
Overall	28.48	24.05(-4.43)	4.97(-23.51)	27.67	27.58(-0.09)

4.3 Evaluation of Spatially Fine-tuned Models

Recently several approaches have emerged to enhance general-purpose models with spatial reasoning capabilities, either by training them on spatial reasoning data (Jia et al., 2025) or by introducing architectural modifications to equip models with latent spatial representations (Fan et al., 2025; Wu et al., 2025). We evaluate three spatially fine-tuned models under the Uniform-50 setting (Tab. 3). Compared with their respective base models, only SpaceQwen exhibits almost no change in performance, whereas both Spatial-MLLM and VLM3R suffer from degradation, particularly in instruction-following ability, leading to an overall drop in performance. This trend is consistent with observations reported in prior work (Yang et al., 2025; Lin et al., 2025): although such models may perform well on some specific spatial reasoning benchmarks, their capabilities do not generalize effectively to other benchmarks and may even impair original abilities. These findings further highlight the challenge posed by our benchmark and its emphasis on comprehensively assessing models' spatial intelligence.

5 Error Analysis

5.1 Error Categorization

We categorize model errors into five non-overlapping, comprehensive types (Fig. 4). Extended explanations are provided in Appendix G.1.

Detailed Grounding Error. Failures in fine-grained perception, including missing or confusing objects, overlooking subtle temporal changes, or misidentifying events at specific timestamps. **ID Mapping Error.** Failures in maintaining consistent identity tracking across frames, often caused by occlusion, rapid motion, or visually similar distractors, leading the model to confuse or mismatch entities over time. **Geometric Reasoning Error.** Mistakes in inferring spatial relations (relative positions or distance, e.g., front/behind, near/far), revealing the model's inability to establish coherent spatial associations across frames. **Prompt Alignment Error.** Misunderstandings in interpreting the prompt or integrating it with visual evidence. **Latent Logical Inference Error.** Failures in reasoning that require integrating implicit cues or commonsense knowledge.

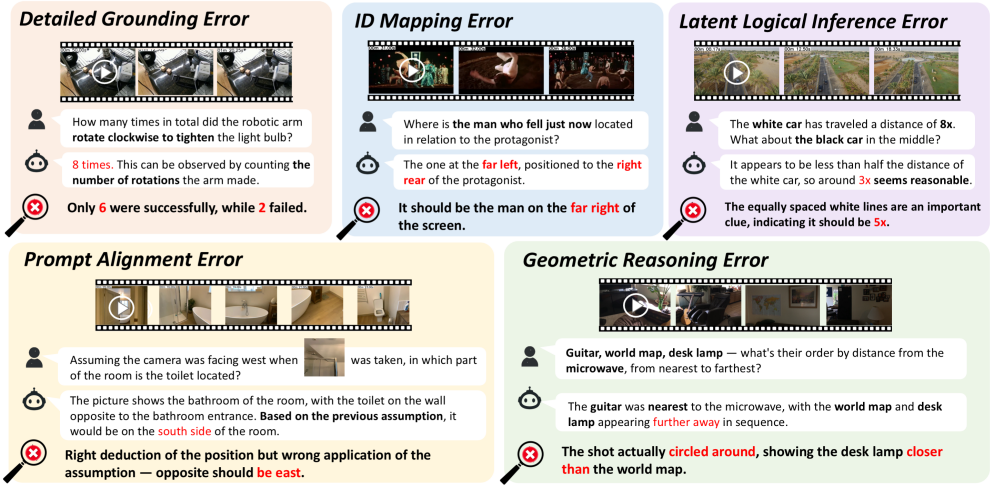

Figure 4: Illustration of five error types, along with examples of model responses and error analyses.

5.2 Error Statistics

We select four representative models (GPT-4o, Gemini 2.5 Flash, O3, and QwenVL2.5-72B) and conducted an error analysis on a total of 520 incorrectly answered cases, evenly sampled across different question categories. The errors were categorized and quantified, and the final statistics, shown in Fig. 5, illustrate the distribution of each error type within the main categories, as well as the overall composition of error types. Several observations can be made:

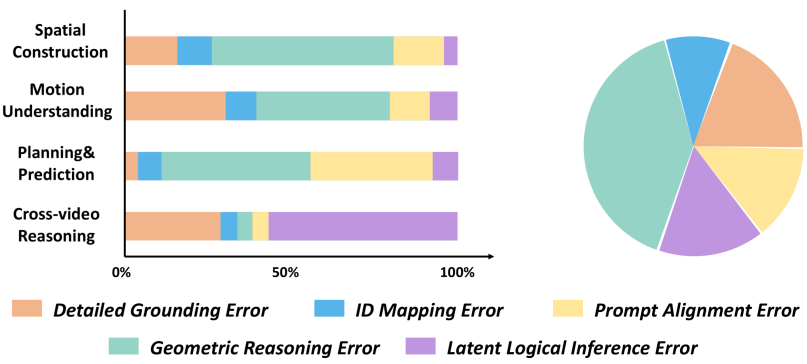

Figure 5: Distribution of the five error types. (Left) Error distribution across different question categories. (Right) Overall proportion of each error type.

(a) Geometric Reasoning Error is the most prevalent error type overall, especially within the **Spatial Construction tasks**. This finding is consistent with prior spatial reasoning benchmarks (Yang et al., 2024; Lin et al., 2025; Yang et al., 2025), indicating that current models still struggle with inferring even simple geometric relations.

(b) Distinct error distributions reveal task-specific capability bottlenecks. Beyond Spatial Construction, we observe: (i) In **Motion Tasks**, detailed grounding errors are prominent. Despite strong single-frame detection and cross-frame ID alignment, models struggle with long-term temporal consistency, suggesting an over-reliance on short-term/local patterns. This points to the need for explicit temporal supervision and temporal state aggregation mechanisms, rather than independent frame-wise encoding.

(ii) In **Prediction & Planning Tasks**, prompt alignment error is prevalent. Even with correct spatial or spatiotemporal understanding, models fail to align with high-level goals, assumptions, or contextual conditions. This indicates that intermediate representations are not optimized for decision-making, motivating instruction-conditioned objectives and explicit mappings from structured states to outputs.

(iii) In **Cross-Video Reasoning Tasks**, latent logical inference errors are the most prominent, followed by detailed grounding errors. These tasks require identifying correspondences across multiple videos, yet models often fail to locate the same instance in both inputs or to leverage such correspondences effectively for reasoning. That exposes weak cross-input relational modeling and motivates explicit cross-video alignment with shared latent representations, rather than simple sequence concatenation.

Through this error analysis, we identify the key failure modes in each MMSI-Video-Bench category and highlight model weaknesses to target, with discussions and preliminary improvement strategies provided in Appendix D.

6 Limitations and Conclusion

We present MMSI-Video-Bench, a diverse, human-annotated, and holistic video-based spatial intelligence benchmark. Our evaluation reveals a substantial gap between model and human performance, with models struggling across all task categories beyond spatial construction, and even spatially fine-tuned models failing to generalize effectively to our bench. Error analyses expose task-specific failure patterns that highlight concrete weaknesses in current models; our preliminary explorations further show that neither 3D spatial cues nor COT prompting yields meaningful gains; and the frame-sampling study underscores the benchmark's difficulty and the need for more effective sampling strategies. Overall, MMSI-Video-Bench provides a rigorous and holistic testbed for assessing spatial intelligence in video models, while offering actionable insights and directions for future improvement. **Limitation.** One potential limitation of MMSI-Video-Bench is its relatively smaller scale compared to other benchmarks, which we discuss in more detail in the Appendix A.5.

AI USE STATEMENT

Generative AI tools were used for language polishing of the manuscript.

REFERENCES

Dl3dv license. https://github.com/DL3DV-10K/Dataset/blob/main/DL3DV-10K_term_of_use.pdf.

Ego4d license. https://ego4d-data.org/pdfs/Ego4D-Licenses-Draft.pdf.

Scannet license. https://kaldir.vc.in.tum.de/scannet/ScanNet_TOS.pdf, a.

Scannet++ license. https://scannetpp.mlsg.cit.tum.de/scannetpp/static/scannetpp-terms-of-use.pdf, b.

Shengnan An, Xunliang Cai, Xuezhi Cao, Xiaoyu Li, Yehao Lin, Junlin Liu, Xinxuan Lv, Dan Ma, Xuanlin Wang, Ziwen Wang, and Shuang Zhou. Amo-bench: Large language models still struggle in high school math competitions, 2025. URL https://arxiv.org/abs/2510.26768.

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, 2025. URL https://arxiv.org/abs/2502.13923.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL https://openreview.net/forum?id=tjZjv_qh_CE.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving, 2020. URL https://arxiv.org/abs/1903.11027.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities, 2024. URL https://arxiv.org/abs/2401.12168.

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, Yiming 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. URL https://arxiv.org/abs/2412.05271.

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 models, 2024. URL https://arxiv.org/abs/2406.01584.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018.

Heng Fan, Hexin Bai, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Harshit, Mingzhen Huang, Juehuan Liu, Yong Xu, Chunyuan Liao, Lin Yuan, and Haibin Ling. Lasot: A high-quality large-scale single object tracking benchmark, 2020. URL https://arxiv.org/abs/2009.03465.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Dilin Wang, Zhicheng Yan, Hongyu Xu, Justin Theiss, Tianlong Chen, Jiachen Li, Zhengzhong Tu, Zhangyang Wang, and Rakesh Ranjan. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction, 2025. URL https://arxiv.org/abs/2505.20279.

Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 653–660. IEEE, 2024.

Chaoyou Fu, Yuhan Dai, Yondong 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*, 2024a.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive, 2024b. URL https://arxiv.org/abs/2404.12390.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video, 2022. URL https://arxiv.org/abs/2110.07058.

Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions, 2018. URL https://arxiv.org/abs/1705.08421.

Mingfei Han, Liang Ma, Kamila Zhumakhanova, Ekaterina Radionova, Jingyi Zhang, Xiaojun Chang, Xiaodan Liang, and Ivan Laptev. Roomtour3d: Geometry-aware video-instruction tuning for embodied navigation, 2025. URL https://arxiv.org/abs/2412.08591.

Ruize Han, Wei Feng, Jiewen Zhao, Zicheng Niu, Yunjun Zhang, Liang Wan, and Song Wang. Complementary-view multiple human tracking. In *AAAI Conference on Artificial Intelligence*, 2020.

Yuping He, Yifei Huang, Guo Chen, Baoqi Pei, Jilan Xu, Tong Lu, and Jiangmiao Pang. Egoexobench: A benchmark for first- and third-person view video understanding in mllms, 2025. URL https://arxiv.org/abs/2507.18342.

Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Dong Lu, Yali Wang, Limin Wang, and Yu Qiao. Egoexolearn: A dataset for bridging asynchronous ego- and exo-centric view of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Rasmus Jensen, Anders Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 406–413. IEEE, 2014.

Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-Chun Zhu. Lemma: A multiview dataset for learning multi-agent multi-view activities. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. *arXiv preprint arXiv:2506.03135*, 2025.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, Vitor Guizilini, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Muhammad Zubair Irshad, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. Droid: A large-scale in-the-wild robot manipulation dataset. 2024.

Anastasia Kirillova, Eugene Lyapustin, Anastasia Antsiferova, and Dmitry Vatolin. Erqa: Edge-restoration quality assessment for video super-resolution. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. SCITEPRESS - Science and Technology Publications, 2022. doi: 10.5220/0010780900003124. URL http://dx.doi.org/10.5220/0010780900003124.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark, 2024. URL https://arxiv.org/abs/2311.17005.

Yifei Li, Junbo Niu, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, and Jiaqi Wang. Ovo-bench: How far is your video-llms from real-world online video understanding?, 2025a. URL https://arxiv.org/abs/2501.05510.

Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13536–13545, October 2021.

Yun Li, Yiming Zhang, Tao Lin, XiangRui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. Stibench: Are mllms ready for precise spatial-temporal world understanding? *arXiv preprint arXiv:2503.23765*, 2025b.

JingLi Lin, Chenming Zhu, Runsen Xu, Xiaohan Mao, Xihui Liu, Tai Wang, and Jiangmiao Pang. Ost-bench: Evaluating the capabilities of mllms in online spatio-temporal scene understanding. *arXiv preprint arXiv:2507.07984*, 2025.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, pp. 445–461, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46448-0.

Matthias Müller, Adel Bibi, Silvio Giancola, Salman Al-Subaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild, 2018. URL https://arxiv.org/abs/1803.10794.

Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip H. S. Torr, and Song Bai. Occluded video instance segmentation: A benchmark, 2022. URL https://arxiv.org/abs/2102.01558.

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ivan Laptev, Ali Farhadi, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. *ArXiv e-prints*, 2016. URL http://arxiv.org/abs/1604.01753.

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 Ettinger, 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)*, June 2020.

Kexian Tang, Junyao Gao, Yanhong Zeng, Haodong Duan, Yanan Sun, Zhening Xing, Wenran Liu, Kaifeng Lyu, and Kai Chen. Lego-puzzles: How good are mllms at multi-step spatial reasoning? *arXiv preprint arXiv:2503.19990*, 2025a.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding, 2025b. URL https://arxiv.org/abs/2502.21271.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024. URL https://arxiv.org/abs/2406.16860.

Hemish Veeraboina. Aime problem set 1983-2024, 2023. URL https://www.kaggle.com/datasets/hemishveeraboina/aime-problem-set-1983-2024.

Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance re-localization in changing indoor environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7658–7667, 2019.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer, 2025a. URL https://arxiv.org/abs/2503.11651.

Shaoguang Wang, Ziyang Chen, Yijie Xu, Weiyu Guo, and Hui Xiong. Less is more: Token-efficient video-qa via adaptive frame-pruning and semantic graph integration, 2025b. URL https://arxiv.org/abs/2508.03337.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024. URL https://arxiv.org/abs/2407.15754.

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 (CVPR)*, pp. 9777–9786, June 2021.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1645–1653, 2017.

Runsen Xu, Weiyao Wang, Hao Tang, Xingyu Chen, Xiaodong Wang, Fu-Jen Chu, Dahua Lin, Matt Feiszli, and Kevin J. Liang. Multi-spatialmllm: Multi-frame spatial understanding with multi-modal large language models. *arXiv preprint arXiv:2505.17015*, 2025.

Jihan Yang, Shusheng Yang, Anjali 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.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025.

Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Rouyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in mllms. *arXiv preprint arXiv:2504.15280*, 2025.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–22, 2023.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9127–9134, 2019.

Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yujie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *arXiv preprint arXiv:2503.22976*, 2025a.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data, 2025b. URL https://arxiv.org/abs/2410.02713.

Ziwei Zhao, Yuchen Wang, and Chuhua Wang. Fusing personal and environmental cues for identification and segmentation of first-person camera wearers in third-person views. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16477–16487, 2024. URL https://api.semanticscholar.org/CorpusID:270358888.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images, 2018. URL https://arxiv.org/abs/1805.09817.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

APPENDIX

A BENCHMARK DETAILS

A.1 TASK FORMULATION DETAILS

As defined in the main paper, MMSI-Video-Bench is organized into five main categories: Spatial Construction, Motion Understanding, Planning, Prediction, and Cross-Video Reasoning. Spatial Construction evaluates the spatial attributes of instances and scenes, as well as the pairwise spatial relations among instances, scenes, and the camera. Motion Understanding is further divided into three aspects: camera motion, instance motion, and inter-instance interactive motions. Cross-Video Reasoning encompasses two subtypes: Memory Update and Multi-View Integration. In total, MMSI-Video-Bench consists of 5 main categories and 13 subtypes, with their detailed definitions summarized in Fig. 6.

Categories	*Subtypes*	*Description*
Spatial Construction	Inst./Scene Spatial Attribute	*Quantity, shape, and positional states of instances or scenes.*
	Inst.–Inst. Spatial Rel.	*Spatial relationships between instances (direction or distance).*
	Scene–Scene Spatial Rel.	*Spatial relationships between scenes.*
	Inst.–Scene Spatial Rel.	*Spatial relationships between instances and the scenes.*
	Cam.–Inst. Spatial Rel.	*Spatial relations between the camera (observer) and instances.*
	Cam.–Scene Spatial Rel.	*Spatial relations between the camera (observer) and scenes.*
Motion Understanding	Cam. Motion	*Motion of the camera, including local movement states and global trajectories.*
	Inst. Motion	*Motion of instances, covering both local movement states and global trajectories.*
	Interactive Motion	*Interaction dynamics among multiple instances, including both local interactions and global patterns.*
Planning	*Plan toward a specific goal based on all available video information.*	
Prediction	*Perform prediction or imagination based on all video information or given prior conditions.*	
Cross-Video Reasoning	Memory Update	*Update knowledge across temporally separated video segments, focusing on reasoning over changes and updates.*
	Multi-View Integration	*Fuse information from videos captured from different viewpoints, focusing on integrative reasoning.*

Figure 6: Task categories and subtypes in MMSI-Video-Bench. "Inst." denotes "instance", "Cam." denotes "camera" and "Rel." denotes "relationship".

A.2 DATA COLLECTION & PREPROCESSING DETAILS

Our data sources include 25 publicly available datasets (Han et al., 2025; Dai et al., 2017; Yeshwanth et al., 2023; Wald et al., 2019; Baruch et al., 2021; Jensen et al., 2014; Fan et al., 2020; Mueller et al., 2016; Müller et al., 2018; Caesar et al., 2020; Damen et al., 2018), such as ScanNet (Dai et al., 2017), RealEstate10k (Zhou et al., 2018), DL3DV (Ling et al., 2024), Ego4D (Grauman et al., 2022), DROID (Khazatsky et al., 2024), Waymo (Sun et al., 2020) and MultiSports (Li et al., 2021) (the remaining datasets are listed in the appendix), amounting to approximately 20k video clips.

Our benchmark is constructed from 25 publicly available video datasets, complemented by additional videos captured and collected by ourselves. During the data preprocessing, we perform filtering to remove clips that are too short in duration, and for datasets originally provided in frame format, we reconstruct videos by concatenating frames according to the FPS specified in their corresponding papers.

Each video dataset falls into one of several capture types, including indoor scanning, outdoor environment, egocentric interactions, exocentric human activities, and other categories. As mentioned in the main paper, we standardize the frame rate for each video category to an appropriate value that ensures no key information is lost. The capture types, frame-rate settings, and average duration statistics for each category are summarized in Tab. 4.

Table 4: Statistics of all source video datasets, including their capture types, average durations, and standardized FPS after pre-processing.

Dataset	Capture Type	Frames Per Second (NUM./Sec.)	Video Duration (Sec.)
Roomtour3d (Han et al., 2025)	Indoor Scanning	1.0	466.86
ScanNet (Dai et al., 2017)	Indoor Scanning	1.0	39.92
ScanNet++ (Yeshwanth et al., 2023)	Indoor Scanning	2.0	136.35
3RScan (Wald et al., 2019)	Indoor Scanning	1.0	60.94
ARKitScenes (Baruch et al., 2021)	Indoor Scanning	1.0	77.28
RealEstate10k (Zhou et al., 2018)	Indoor Scanning	0.66	214.73
DL3DV (Ling et al., 2024)	Indoor&Outdoor Environment	1.0	39.81
Waymo (Sun et al., 2020)	Outdoor Environment	5.0	15.92
NuScenes (Caesar et al., 2020)	Outdoor Environment	4.0	26.79
OVIS (Qi et al., 2022)	Outdoor Environment	5.0	13.74
TrackingNet (Müller et al., 2018)	Outdoor Environment	4.0	21.33
LaSOT (Fan et al., 2020)	Outdoor Environment	5.89	32.40
UAV123 (Mueller et al., 2016)	Outdoor Environment	5.89	24.32
Ego4D (Grauman et al., 2022)	Egocentric Interactions	2.0/8.33	262.49
EPIC-KITCHENS (Damen et al., 2018)	Egocentric Interactions	2.0/8.33	91.51
EgoExoLearn (Huang et al., 2024)	Egocentric Interactions	4.0	565.81
MultiSports (Li et al., 2021)	Exocentric Human Activities	2.0/8.33	20.51
charades (Sigurdsson et al., 2016)	Exocentric Human Activities	4.0	27.91
LEMMA (Jia et al., 2020)	Exocentric Human Activities	12.50	22.84
TF2023 (Zhao et al., 2024)	Exocentric Human Activities	2.0	492.46
CVMHT (Han et al., 2020)	Exocentric Human Activities	4.0	37.25
AVA (Gu et al., 2018)	Exocentric Human Activities	1.0	900.27
DROID (Khazatsky et al., 2024)	Others	4.0	86.75
RH20T (Fang et al., 2024)	Others	4.0	84.32
DTU (Jensen et al., 2014)	Others	2.0	24.00
RealWorld	Indoor&Outdoor Environment	2.0	46.43

A.3 Human Annotation UI

As shown in Fig. 7, we provide a dedicated UI tool for both annotation and validation. The interface allows users to switch between Annotation Mode and Validation Mode. In the annotation mode, annotators can select a question type, choose the corresponding video, and determine appropriate start/end frames to construct a question. The system supports questions in either pure text or text+image format. Annotators then design the answer options, assign the correct answer, and provide the reasoning behind it. The UI clearly displays timestamps corresponding to each frame, helping annotators position temporal cues precisely. Additionally, to improve annotation efficiency, we provide a Video Browsing Assistant Tool, enabling quick coarse-level navigation and preview of video content.

In the validation mode, validators are randomly assigned samples annotated by others. They can inspect the loaded annotation and choose to Accept or Reject it. For rejected samples, the validator is required to provide reasons and suggestions for revision.

A.4 Data Statistics

In Fig. 9, we present the distribution of video durations. And Fig. 8 illustrates a word cloud of the benchmark annotations, together with the distribution of video source capture types.

A.5 Benchmark Scale Considerations

Although MMSI-Video-Bench is not large in terms of data scale, with approximately 1k annotated samples, this scale is reasonable when considering the following factors:

Rationale for Our Human-Curated, Compact, and Challenging Spatial Benchmark. Our benchmark is a high-difficulty spatial intelligence benchmark fully designed and annotated by hu-

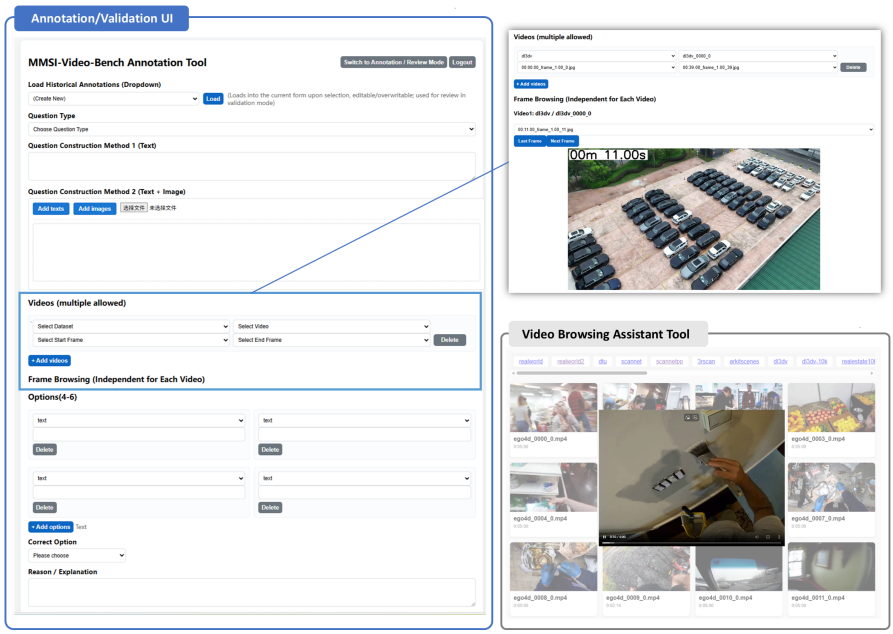

Figure 7: User interface for annotation and quality validation in MMSI-Video-Bench.

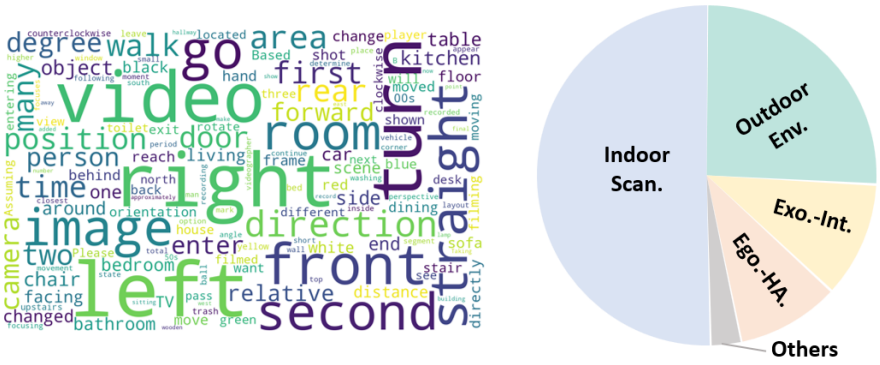

Figure 8: (Left) Word cloud of MMSI-Video-Bench. (Right) Distribution of video source categories across all samples in MMSI-Video-Bench.

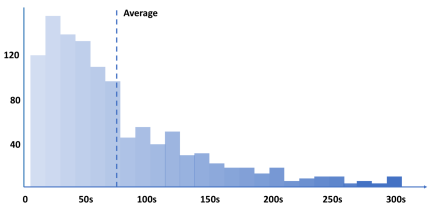

Figure 9: Duration distribution of all video samples in MMSI-Video-Bench (in seconds).

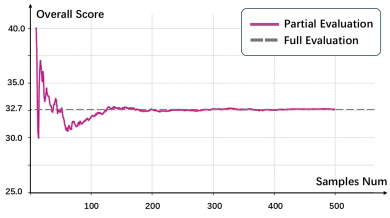

Figure 10: Trend of overall score of QwenVL2.5-72B as the number of evaluated samples increases.

man researchers. Its primary goal is not to cover a large sample distribution, but to systematically probe the upper bounds and critical failure modes of current models in spatial intelligence. Specifically, all 1,106 questions are carefully designed by expert researchers and undergo a rigorous manual annotation and cross-review process, making it infeasible to obtain easy samples. These questions focus on highly challenging spatial reasoning scenarios and can be regarded as a human-level stress test for spatial intelligence. As shown in Tab. 1, our benchmark consistently exhibits a substantial human–AI performance gap, indicating strong discriminative power in characterizing model capabilities.

In the context of high-difficulty reasoning benchmarks, dataset size is neither the sole nor the primary criterion. There is substantial precedent showing that small-scale but high-quality and highly challenging benchmarks are valuable for evaluating model capabilities (Veeraboina, 2023; An et al., 2025; Yang et al., 2025), they often contain only a few hundred (less than 1,000) samples. To constitute a representative gold benchmark, 1,000 carefully designed high-difficulty samples are sufficient.

We additionally conduct an experiment to demonstrate that the scale of our benchmark provides stable evaluation results. Specifically, we randomly choose 10 to 500 instances from the full set of 1,106 samples and perform partial evaluation, reporting the performance of QwenVL2.5-72B on each sampled set. As shown in Figure 10, the pink curve illustrates the trend of the overall score as the number of evaluated samples increases, while the gray dashed line denotes full evaluation, i.e., the average score computed over all 1,106 samples as a reference. We observe that the overall score gradually converges as the sample size increases and becomes stable once the number of samples exceeds 200, with fluctuations within 2%. These results indicate that a benchmark scale of 1,106 samples is sufficient to provide stable and reliable evaluation. Further analysis of evaluation stability is provided in Appendix A.7.

Scalable Automatic Annotation Is Unsuitable for Constructing Our Benchmark. Currently prevalent scalable data construction approaches, such as template-based or LLM-based annotation, are not suitable for the objectives of our benchmark.

(1) First, although template-based data construction can increase surface-level linguistic diversity through template expansion or refinement using LLMs, its semantic space remains constrained by predefined problem-generation rules. Such methods fundamentally struggle to capture the complexity of real-world spatial reasoning and are vulnerable to rule-based or pattern-matching attacks. High performance on these benchmarks does not necessarily reflect genuine spatial understanding or reasoning ability. For example, some models (e.g., Spatial MLLM (Wu et al., 2025) and VLM3R (Fan et al., 2025)) achieve strong performance on certain benchmarks by scaling up training data constructed with similar rule-based patterns, but their performance drops significantly on other spatial intelligence benchmarks (e.g., MMSI-Bench (Yang et al., 2025), OST-Bench (Lin et al., 2025), and our MMSI-Video-Bench), revealing limited generalization. Moreover, data constructed following fixed rules may introduce implicit shortcuts, further weakening the benchmark's discriminative power.

(2) Second, LLM-based annotation is more suitable for constructing semantic tasks that fall within the model's existing capability range. Our benchmark, however, is designed to probe the limits of spatial reasoning—a domain where existing VLMs and LLMs are still notably weak. Relying on LLMs for annotation in this context is therefore impractical. Moreover, LLM-based annotations may introduce systematic biases, further undermining the reliability of the benchmark.

Overall, we deliberately adopt carefully controlled human design and annotation procedures to ensure the benchmark's challenge, reliability, and long-term value, rather than prioritizing scalability in sample size.

A.6 Benchmark Diversity Analysis

We characterize the diversity of MMSI-Video-Bench along three complementary dimensions: (a) video sources and capture types, (b) subtask design and capability coverage across dataset–subcategory combinations, and (c) question formulations and linguistic expression. Together, these dimensions describe both what the benchmark evaluates and how its questions are expressed.

Diversity of video sources and capture types. Previous spatial reasoning benchmarks are typically limited to a small number of capture settings or narrowly defined domains. In contrast, to better

reflect the complexity of real-world spatial intelligence, MMSI-Video-Bench integrates videos from 25 public datasets and in-house recordings, counted as 26 sources, covering a wide spectrum of scenarios, including tabletop settings, indoor environments (ranging from single rooms to multi-floor buildings), outdoor scenes such as urban architecture and street views, natural landscapes, sports activities, and movie or TV footage (Tab. 4). This broad coverage leads to a significantly higher diversity of data sources compared to prior benchmarks, as also reflected by the Sources column in Tab. 1.

Diversity of subtask design and evaluated capabilities. Capability diversity arises from differentiated question design and evaluation objectives across 131 dataset–subcategory combinations within the 13 subcategories. For each combination, question construction, reasoning requirements, and evaluation focuses are tailored to the video setting and subtask. All questions are manually designed and annotated by experts, enabling flexible question formulations around the spatial relations, motion, and contextual evidence in each video. This process creates varied spatial reasoning challenges across different types of video data.

Diversity of sample design and expression. As a fully human-annotated benchmark, MMSI-Video-Bench achieves strong diversity at the sample level, reflected in both question design paradigms and linguistic expression. Template-based benchmarks inherently constrain both the diversity of question formats and textual expressions. While LLM-based approaches can partially improve linguistic diversity, their question design is still largely bounded by predefined rules or generation patterns. In contrast, human annotation enables more flexible, non-templated question formulations, leading to richer and more varied samples. As shown in Tab. 5, we further compare textual diversity with other spatial reasoning benchmarks using three metrics: the number of distinct words, unique trigrams, and unique 5-grams. MMSI-Video-Bench has higher counts across these metrics than the listed comparisons. These corpus-level statistics describe the range of language used in the questions and complement the task-design analysis above.

Overall, diverse real-world sources, differentiated subtask designs, and flexible expert-written questions provide complementary forms of coverage. This design supports a broad diagnostic evaluation, while the separate experiments in Appendix A.7 assess score and ranking stability at the full benchmark scale.

Table 5: Linguistic diversity statistics across benchmarks.

Metric	**CV-Bench** (Tong et al., 2024)	**LEGO-Puzzle** (Tang et al., 2025a)	**All-Angles-Bench** (Yeh et al., 2025)	**VSI-Bench** (Yang et al., 2024)	**MMSI-Video-Bench**
Distinct Words	529	259	644	376	2248
Unique Trigrams	3581	973	6381	6528	19580
Unique 5-grams	6655	1264	13225	12395	26020

A.7 Benchmark Stability Analysis

We evaluate the statistical stability of MMSI-Video-Bench at its full scale of 1,106 questions from two complementary perspectives: consistency of scores across repeated evaluations and consistency of model orderings under question resampling.

A.7.1 Score Stability across Repeated Evaluations

We conduct five evaluations on all 1,106 questions for GPT-4o, O3, Gemini 2.5 Flash, and GPT-5. For each model m, we report the range $R_m = \max_{1\le r\le 5} S_{m,r} - \min_{1\le r\le 5} S_{m,r}$, where $S_{m,r}$ is the overall benchmark score, expressed as a percentage, in evaluation r. Table 6 reports the resulting ranges in percentage points.

Table 6: Overall score variation across five evaluations on the full set of 1,106 questions. Range is the maximum score minus the minimum score, in percentage points.

Model	Range (percentage points)
GPT-4o	0.89
O3	1.56
Gemini 2.5 Flash	1.79
GPT-5	1.23

The five-run score ranges are 0.89–1.79 percentage points across the four models, indicating consistent overall scores across repeated evaluations at the full benchmark scale.

A.7.2 Ranking Stability under Bootstrap Resampling

To assess sensitivity of model orderings to the sampled questions, we use a paired, non-parametric bootstrap. Each resample draws 1,106 question indices with replacement from the full benchmark. We compute the compared models' scores on the same resampled questions and repeat this procedure 1,000 times. For models A and B, we estimate the fraction of resamples in which A outperforms B:

$$\widehat{P}(A > B) = \frac{1}{1000} \sum_{b=1}^{1000} \mathbf{1}\left[S_A^{(b)} > S_B^{(b)}\right], \tag{1}$$

where $S_A^{(b)}$ and $S_B^{(b)}$ are their overall scores in bootstrap resample b.

The observed ordering of the four representative models is Gemini 3 Pro (m_1), Gemini 2.5 Flash (m_2), GPT-4o (m_3), and QwenVL2.5-32B (m_4), from highest to lowest accuracy. Table 7 reports support for the three adjacent orderings.

Table 7: Bootstrap support for adjacent model orderings over 1,000 resamples of 1,106 questions.

Ordered model pair	$\widehat{P}(A > B)$
Gemini 3 Pro > Gemini 2.5 Flash	0.915
Gemini 2.5 Flash > GPT-4o	0.974
GPT-4o > QwenVL2.5-32B	0.955

Each of the three adjacent model orderings is preserved in more than 90% of bootstrap resamples. Together with the five-run score ranges, these results provide empirical support for score stability and relative ranking stability for the evaluated models at the full scale of 1,106 questions.

B EXPERIMENT DETAILS

B.1 MODEL PERFORMANCE EVALUATION

In our evaluation, all models are provided with the same input template. As illustrated in Fig. 11, the system prompt specifies the timestamp information for each frame and enforces the required output format. The user message injects, in order, the video, a brief task description, and the question with its options into the template. The expected output format adapts to the evaluation setting (e.g. During error analysis we require models to produce both an answer and a reason to facilitate failure localization, while under Chain-of-Thought prompting we require the model to emit each intermediate thinking step in addition to the final answer.)

For model outputs, we employ a general-purpose answer extraction function to parse the predicted answers. This ensures consistent extraction across different models, and we verify that all correct responses produced by any model can be accurately detected and extracted.

System Prompt	*Please answer questions based on video content. When processing each query:* *(1) Each video is presented as a frame sequence, with each frame's top-left corner displaying a timestamp ("-- m -- s") that indicates the exact capture time. The timestamps on the initial and final frames correspond to the video's start and end times.* *(2) Please format your response exactly as specified in the question requirements.*
User Message	**<video>** **<task brief description>** e.g. This task focuses on the interactions between instances. Next, I will ask a question based on the video information. My question is: **<question>.** Please select the correct answer from the following options: **<options>** A. You only need to output a single letter (**<letters>**) as your response. B. You need to provide your detailed reasoning process. Provide your output(answer and reason) in JSON format: {"answer": a letter, "reason": str}.

Figure 11: Structure of system and user prompts used in the experiments.

B.2 HUMAN PERFORMANCE EVALUATION

For our human performance evaluation, we recruited six participants with a certain level of education, including students and members of the general public. None of them had prior in-depth knowledge or experience in this domain; they were only given instructions on how to answer the questions. It took them approximately five days to complete all 1106 questions. We evaluated their responses and averaged the results to obtain the final human performance metric.

C MORE ANALYSIS

C.1 MORE FINDINGS OF THE MAIN RESULTS

Prediction is the most challenging main category, and Camera–Instance Spatial Relation is the most challenging subtype. In Tab. 2, performance on Prediction is generally lower than other main task categories (i.e., the average scores of Spatial Construction, Motion Understanding, Planning, and Cross-Video Reasoning). This is because these tasks require models to go beyond simply understanding the spatio-temporal information and instead make predictions based on specific conditions or physical priors. Among the various types of Spatial Construction subtasks, the Camera–Instance Spatial Relation subtype is the most challenging. This is due to its combination of ego-to-scene spatial reasoning and detailed grounding of instances within the video, making it the most difficult among all subtypes.

Model performance across main categories and difficulty Levels. We further computed the average scores of each model across the main categories, as shown in Tab. 8. Overall, Gemini 3 Pro achieves the highest performance, while GPT-5 demonstrates the strongest capabilities in spatial construction and reasoning tasks. On the other hand, Seed-1.6-vision excels at motion understanding. As a model with high reasoning and decision-making capacity, Gemini 3 Pro also leads

in Prediction and Planning tasks compared to other models. Meanwhile, O3 excels in reasoning across multiple videos. Furthermore, we categorize our benchmark into three difficulty levels: easy, medium, and hard, based on the overall accuracy of all models on each question. Given our categorization criteria, it is natural that the scores follow the trend: hard < medium < easy. Examining model performance across these difficulty levels, we find that Gemini 3 Pro and Gemini 2.5 Flash are particularly effective at solving questions that most other models struggle with.

Table 8: Model performance across main categories and difficulty levels. "SC." abbrev for "Spatial Construction", "MU." abbrev for "Motion Understanding" and "CV." abbrev for "Cross-Video Reasoning".

Methods	Avg	SC.	MU.	Plan.&Pred.	CV.	Hard	Medium	Easy
(Sufficient-Coverage) Proprietary								
O4-mini	35.1	34.9	35.2	34.0	36.6	16.2	30.0	**56.7**
O3	**37.3**	39.0	35.6	33.5	**40.1**	19.3	35.0	55.4
GPT-4o	28.1	28.0	26.1	28.6	30.8	13.2	26.0	43.3
Gemini 2.5 Flash	36.6	**39.4**	37.1	31.6	34.3	**21.7**	32.8	53.6
-Thinking	36.7	36.6	**38.3**	**34.5**	37.2	21.1	**35.2**	51.7
(Sufficient-Coverage) Open-source								
InternVL2.5-8B	28.7	27.4	31.1	27.7	29.6	13.2	24.5	46.4
InternVL3-8B	29.6	28.7	33.0	25.2	32.0	**15.6**	26.0	45.4
InternVideo2.5-8B	26.9	25.9	25.4	29.1	29.1	14.4	21.2	43.5
QwenVL2.5-7B	28.8	25.4	34.1	30.1	28.5	13.8	26.5	44.3
QwenVL2.5-32B	**32.4**	**31.0**	**35.2**	32.0	32.0	13.8	**31.0**	49.9
QwenVL2.5-72B	31.8	28.0	34.9	**33.5**	**35.5**	13.8	24.8	**54.9**
QwenVL3-8B	29.1	27.4	32.3	28.7	29.0	12.8	22.0	51.0
QwenVL3-30B	29.1	28.4	27.6	32.5	29.1	12.2	22.0	51.2
-Thinking	28.0	30.4	26.9	26.2	25.6	11.9	23.0	47.2
(Uniform-50) Proprietary								
Claude-haiku-4.5	34.3	33.0	37.5	32.0	**35.5**	18.0	30.5	52.2
GPT-5	36.8	**41.4**	36.0	32.0	31.4	20.2	32.5	55.7
O4-mini	34.2	37.1	35.6	28.2	31.4	15.0	29.0	**56.2**
O3	37.0	38.4	37.1	35.4	34.9	20.5	34.2	54.1
GPT-4o	31.6	31.2	33.3	31.6	29.6	17.7	29.2	45.9
Gemini 3 Pro	**38.0**	39.2	37.9	**37.4**	**35.5**	**22.9**	35.2	53.8
Gemini 2.5 Flash	35.4	39.0	36.4	28.2	33.1	19.0	32.5	52.8
-Thinking	35.2	36.2	35.2	36.9	30.2	19.3	33.0	51.2
Doubao-1.5-thinking	31.6	31.9	34.9	27.1	31.2	13.2	31.9	48.1
Seed-1.6-vision	34.9	33.5	**38.9**	32.6	34.9	16.0	**35.3**	51.5
(Uniform-50) Open-source								
InternVL2.5-8B	29.1	28.0	30.3	27.2	32.6	13.8	23.2	48.5
InternVL2.5-38B	31.0	26.1	34.9	**35.9**	32.6	11.9	27.5	51.2
InternVL2.5-78B	31.4	31.7	31.1	32.0	30.2	14.7	25.0	52.5
InternVL3-8B	30.4	26.9	35.2	30.6	32.0	13.8	27.5	47.8
InternVL3-38B	28.8	30.0	29.2	25.7	29.1	9.2	28.0	46.7
InternVL3-78B	32.5	**32.8**	34.9	26.2	36.0	14.4	27.8	53.3
InternVideo2.5-8B	27.4	26.3	27.3	27.7	30.2	15.3	23.0	42.5
LLaVA-Video-7B	28.5	26.5	32.6	26.7	29.6	**15.9**	27.0	40.9
LLaVA-Video-72B	30.4	27.6	36.4	29.1	30.2	14.1	25.8	49.3
QwenVL2.5-7B	29.7	28.9	33.0	32.0	23.8	11.3	**29.0**	46.2
QwenVL2.5-32B	28.6	25.9	31.8	29.1	30.2	13.2	22.5	48.3
QwenVL2.5-72B	**32.7**	29.3	**38.6**	28.2	**38.4**	11.9	26.5	**57.3**
QwenVL3-8B	27.6	27.4	29.2	26.7	26.7	8.0	21.8	50.7
QwenVL3-30B	28.8	26.5	29.9	30.1	31.4	8.3	26.2	49.1
-Thinking	30.8	31.7	27.3	29.6	35.5	11.9	27.5	50.7

C.2 Frame Sampling Study

Effect of Frame Count and Sampling Strategy. We evaluated model performance with varying frame counts (1, 10, and 50) and two sampling strategies: consecutive frames from a local segment versus uniformly sampled frames across the entire video. Experiments were conducted on three representative models (GPT-4o, Gemini 2.5 Flash, and QwenVL2.5-72B), with results shown as six curves in Fig. 12. The results reveal two key findings: (1) performance is very low at minimal frame counts, sometimes near random guessing level, indicating that there are no shortcuts in MMSI-Video-Bench; performance improves significantly as more frames are sampled, showing the necessity of visual information in MMSI-Video-Bench. (2) Uniform sampling substantially outper-

forms consecutive sampling, demonstrating that broad temporal coverage is essential to capture key events, and that short continuous segments are insufficient. This underscores that MMSI-Video-Bench is designed to require models to integrate information across the full temporal span of the video.

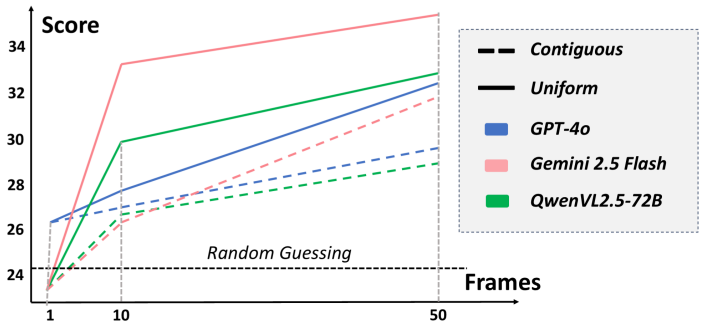

Figure 12: Model performance under different frame sampling methods. Dashed lines indicate contiguous sampling, while solid lines indicate uniform sampling.

Smarter Keyframe Sampling Strategy. Recent studies have proposed more efficient frame sampling strategies that can yield notable performance improvements. Following the Adaptive Keyframe Sampling (AKS) approach (Tang et al., 2025b), which selects frames based on image–text semantic representations, we sampled 50 frames per video and evaluated model performance. Results are summarized in Table 9. Although AKS achieves substantial gains on benchmarks such as LongVideoBench (Wu et al., 2024) and Video-MME (Fu et al., 2024a), it fails to provide improvements on MMSI-Video-Bench. This may be because the key frames required to answer questions in MMSI-Video-Bench cannot be directly determined from semantic similarity alone (e.g., "How does the dog move during the period when it is out of my sight?"). Relying solely on semantic cues may even narrow the model's effective field of view, causing it to miss other critical frames. This outcome underscores the challenging nature of our bench and indicates that it places stricter requirements on frame sampling strategies than existing video benchmarks.

Table 9: Comparison of model performance under uniform sampling and AKS sampling strategies.

Method	GPT-4o	QwenVL2.5-72B	InternVL3-78B
Uniform-50	31.6	32.7	32.5
AKS-50	28.4 (-3.2)	31.9 (-0.8)	31.6 (-0.9)

C.3 Additional Perspectives of MMSI-Video-Bench

Due to the diversity of data sources and the holistic coverage of task types in MMSI-Video-Bench, the benchmark can be also examined from several domain-oriented perspectives. Based on different application focuses, we further derive three subset benchmarks from MMSI-Video-Bench and report model performance on each of them. Similarly, model performance is evaluated under both the Uniform-50 and Sufficient-Coverage settings.

Indoor Scene Perception Bench. The Indoor Scene Perception Bench focuses on evaluating a model's ability to perceive and understand indoor environments. This subset contains 523 samples from MMSI-Video-Bench and includes three major categories: Static-Scene (Instance-Centric), Static-Scene (Camera-Centric), and Dynamic-Scene. The two static-scene categories assess a model's understanding of static indoor layouts. The Instance-Centric category includes questions that are independent of the camera or viewer perspective, targeting object-intrinsic spatial attributes and inter-object spatial relations within the scene. In contrast, the Camera-Centric category examines spatial relations defined relative to the viewer or camera, evaluating the model's understanding of its positional relationship to the surrounding environment. The Dynamic-Scene category tests a

Table 10: Performance of different models on the three subset benchmarks, including overall results and scores for each task subtype. "IC" and "CC" denote Instance-Centric and Camera-Centric, respectively; "Man." and "Nav." represent Manipulation and Navigation; and "TG" and "TL" refer to Target Grounding and Time Localization, respectively.

Model	Indoor Scene Perception				Robot			Grounding		
	Avg.	Static-CC.	Static-IC.	Dynamic	Avg.	Man.	Nav.	Avg.	TG.	TL.
(Sufficient-Coverage) Proprietary										
O4-mini	37.1	38.2	33.5	**49.2**	35.3	28.9	**41.1**	31.3	32.4	25.9
O3	**39.4**	37.6	38.2	**49.2**	36.3	**38.1**	34.6	37.6	38.1	35.2
GPT-4o	29.6	28.0	29.0	36.9	28.9	25.8	31.8	29.2	30.2	24.1
Gemini 2.5 Flash	**39.4**	**39.8**	**39.0**	40.0	35.8	**38.1**	33.6	37.6	**39.9**	25.9
-Thinking	37.9	36.6	37.1	44.6	**37.2**	35.0	39.2	**38.8**	38.8	**38.9**
(Sufficient-Coverage) Open-source										
InternVL2.5-8B	28.7	26.3	27.9	**38.5**	27.9	33.0	23.4	30.1	29.5	33.3
InternVL3-8B	27.7	24.2	30.5	26.1	29.9	33.0	27.1	30.8	**30.6**	31.5
InternVideo2.5-8B	26.8	24.2	26.5	35.4	27.9	25.8	29.9	27.8	28.5	24.1
QwenVL2.5-7B	24.5	24.2	24.6	24.6	26.5	29.9	23.4	28.7	28.1	31.5
QwenVL2.5-32B	29.6	**30.1**	30.5	24.6	32.8	32.0	33.6	**31.0**	29.9	**37.0**
QwenVL2.5-72B	29.2	26.3	29.0	**38.5**	**37.8**	40.2	**35.5**	30.4	**30.6**	29.6
QwenVL3-8B	27.9	24.3	29.6	31.1	32.1	**40.2**	24.8	26.5	26.8	25.0
QwenVL3-30B	**30.0**	29.0	29.0	36.9	32.8	33.0	32.7	28.1	28.1	27.8
-Thinking	29.8	26.9	**32.7**	26.1	27.9	32.0	24.3	26.9	25.3	35.2
(Uniform-50) Proprietary										
Claude-haiku-4.5	33.5	34.4	32.4	35.4	34.8	39.2	30.8	32.8	32.0	37.0
GPT-5	**41.7**	40.3	**42.6**	41.5	37.8	39.2	36.5	35.2	35.6	33.3
O4-mini	37.5	40.3	36.0	35.4	33.3	37.1	29.9	34.3	33.5	**38.9**
O3	40.7	43.0	38.2	**44.6**	39.2	36.1	**42.1**	37.3	37.7	35.2
GPT-4o	31.7	38.2	26.5	35.4	29.9	28.9	30.8	31.9	31.3	35.2
Gemini 3 Pro	39.4	36.0	40.8	43.1	**40.2**	38.1	**42.1**	35.2	35.6	33.3
Gemini 2.5 Flash	39.2	**44.6**	36.8	33.9	33.8	38.1	29.9	**38.2**	**38.1**	**38.9**
-Thinking	36.7	39.8	33.8	40.0	39.7	39.2	40.2	36.1	36.3	35.2
Doubao-1.5-thinking	33.0	30.8	32.5	41.9	36.1	**41.4**	31.2	37.0	37.1	36.8
Seed-1.6-vision	34.2	36.1	31.9	37.2	39.3	**41.4**	37.5	33.0	33.3	31.6
(Uniform-50) Open-source										
InternVL2.5-8B	29.4	25.8	30.1	36.9	28.4	28.9	28.0	28.4	28.8	25.9
InternVL2.5-38B	28.3	24.7	27.9	**40.0**	36.3	36.1	**36.5**	31.9	31.3	35.2
InternVL2.5-78B	30.4	28.5	31.6	30.8	34.8	33.0	**36.5**	29.9	29.2	33.3
InternVL3-8B	27.0	22.0	29.4	30.8	**37.8**	39.2	**36.5**	31.9	31.0	**37.0**
InternVL3-38B	29.1	24.7	31.6	30.8	27.9	28.9	27.1	30.4	31.0	27.8
InternVL3-78B	**32.5**	28.5	**34.6**	35.4	34.3	38.1	30.8	**35.5**	**35.2**	**37.0**
InternVideo2.5-8B	26.8	26.9	26.5	27.7	29.9	28.9	30.8	27.2	25.6	35.2
LLaVA-Video-7B	27.5	22.6	28.3	38.5	24.5	29.9	19.6	27.2	27.1	27.8
LLaVA-Video-72B	28.1	20.4	32.0	33.9	34.3	39.2	29.9	31.0	31.3	29.6
QwenVL2.5-7B	27.1	26.9	28.7	21.5	34.8	35.0	34.6	26.6	25.6	31.5
QwenVL2.5-32B	26.6	24.2	26.1	35.4	30.4	33.0	28.0	27.5	26.0	35.2
QwenVL2.5-72B	30.8	29.0	29.8	**40.0**	34.8	**46.4**	24.3	34.3	34.2	35.2
QwenVL3-8B	28.7	27.4	27.6	36.9	27.0	30.9	23.4	28.7	28.1	31.5
QwenVL3-30B	27.5	23.1	29.4	32.3	32.8	35.0	30.8	29.2	28.1	35.2
-Thinking	32.3	**29.6**	33.5	35.4	27.9	29.9	26.2	31.6	30.6	**37.0**

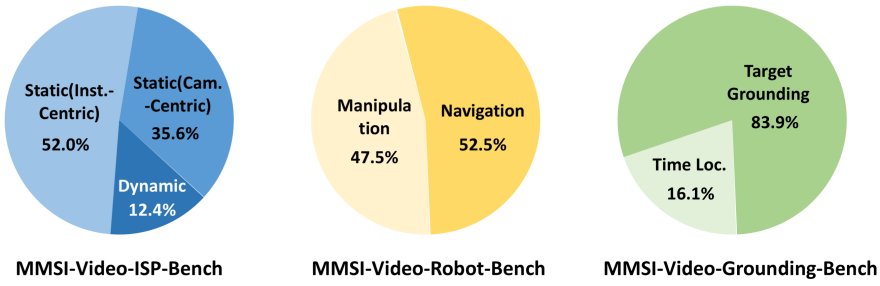

Figure 13: Distribution of subtask proportions across the Indoor Scene Perception Bench, Robot Bench and Grounding Bench.

model's ability to reason about scene changes over time, including those caused by human activities as well as object replacement events that occur between temporally separated video segments.

The evaluation results, shown in the left four columns of Tab. 10, indicate that among all models, GPT-5 achieves the strongest performance; notably, GPT-5 excels in instance-centric static scene perception, while Gemini 2.5 Flash achieves the best performance in camera-centric static scene perception. In addition, O3 and O4-mini show a strong capability in understanding scene changes. Most open-source models lag behind the proprietary ones. Looking across sub-tasks, models with strong overall performance tend to maintain balanced scores across all types, whereas weaker models exhibit their primary bottleneck in Static-CC, which requires reasoning about the spatial relationship between the observer and the environment—a capability that these models struggle with.

Robot Bench. The Robot Bench focuses on evaluating model performance on two core tasks in real-world embodied scenarios: Manipulation and Navigation. This subset contains 204 samples. The Manipulation category assesses a model's ability to perceive and reason about fine-grained tabletop operations and interactive motions, while the Navigation category evaluates a model's planning and navigation capabilities within indoor environments.

As reported in the middle three columns of Tab. 10, Gemini 3 Pro stands out with the strongest overall results on this benchmark. Performance on individual subtasks: QwenVL2.5-72B delivers the best results on Manipulation, while O3 and Gemini 3 Pro lead on Navigation. Notably, the Navigation task reveals substantially larger performance gaps between models and highlights a key weakness of many open-source models.

Grounding Bench. The Grounding Bench comprises 335 samples and requires models to localize either target objects or specific time points within a video. Unlike traditional visual grounding or temporal localization benchmarks that mainly involve semantic referential grounding, our benchmark distinguishes itself by requiring spatial reasoning for all queries to correctly identify the target object or temporal segment. Naturally, this subset is divided into two components based on the type of grounding: target grounding and temporal localization. Within the Grounding Bench, Gemini 2.5 Flash achieves the strongest overall performance, excelling in both temporal localization and target-object identification. O4-mini demonstrates similarly strong temporal localization capabilities as well.

These three benchmarks evaluate model performance within more fine-grained domains and categories, enabling targeted assessment of specific model capabilities. They also provide a convenient evaluation protocol for models designed for particular domains or task types.

D Preliminary Exploration for Model Improvement

While our error analysis categorizes and quantifies the types of failures made by existing models, in this section, we conduct an initial exploration toward improving model performance based on these identified errors.

D.1 Equipping Models with 3D Spatial Cues

Among all error types, Geometric Reasoning Error stems from the model's insufficient ability to build and utilize spatial representations. Correspondingly, we consider two general directions for improving spatial reasoning: (1) training models with sufficient and diverse spatial reasoning data, and (2) enhancing models by explicitly providing or modeling spatial representations, e.g., through dedicated architectures or auxiliary tools. In our preliminary attempt, we adopt the second strategy. Specifically, we equip the model with spatial cues generated by VGGT (Wang et al., 2025a), enabling it to better perceive global scene geometry. As illustrated in Fig. 14, we first feed raw video frames into VGGT to obtain a 3D reconstruction of the scene. We then render 10 multi-view observations (including top-down and multiple side views) from the reconstructed point cloud. These sparse geometric cues are combined with the original video frames and fed into the model together as input.

We evaluate four representative models: Gemini 2.5 Flash, O3, GPT-4o, and QwenVL2.5-72B. Under the Uniform-50 setting, 50 frames from each video were fed into VGGT for 3D reconstruction and rendered into corresponding images. After equipping the models with 3D spatial cues, their performance is shown in Fig. 15. All four models show no significant improvement (with gains

below 1%), suggesting that 3D spatial cues do not reliably enhance spatial intelligence under our current setup. Upon further analysis of model errors, we identified two main issues:

- Issues in generating 3D spatial cues. While VGGT is capable of handling relatively simple scenarios, such as single-room indoor scans, it often breaks down in more complex settings, including multi-room or multi-floor environments, as well as dynamic scenes. In these cases, the rendered views frequently suffer from severe artifacts, incomplete geometry, or incorrect camera alignment, resulting in images that provide little meaningful spatial information and may even introduce misleading cues. These failures indicate inherent limitations of VGGT in reliably modeling complex 3D scene structures. Consequently, more robust and generalizable 3D perception tools are required to consistently produce accurate spatial cues across diverse real-world scenarios.
- Issues in utilizing 3D spatial cues. Beyond the quality of the cues themselves, our analysis of the models' reasoning processes reveals a more fundamental limitation: current MLLMs struggle to effectively utilize 3D spatial cues even when they are provided. In many cases, the cues are partially ignored or incorrectly associated with the video content and the question, despite explicit prompt instructions encouraging their use. This suggests a substantial gap between current MLLMs and genuine 3D spatial understanding.

  At an architectural level, most MLLMs rely on 2D visual backbones and token-based fusion mechanisms that are primarily optimized for appearance-level recognition and semantic alignment, rather than explicit geometric reasoning. Such architectures lack inductive biases for encoding metric structure, spatial continuity, and viewpoint transformations, making it difficult to naturally represent and manipulate 3D relationships. As a result, geometric cues are often treated as auxiliary textual or visual hints instead of being integrated into the core reasoning process. Bridging this gap likely requires not only larger-scale and more diverse training data for spatial alignment, but also architectural designs that are explicitly tailored to modeling 3D structure and spatial relations.

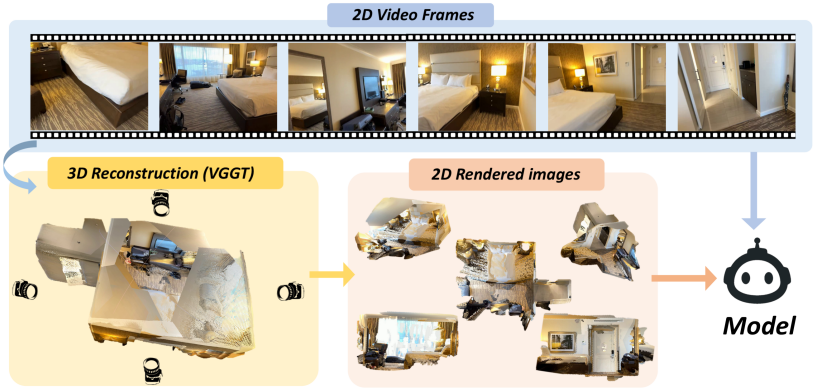

Figure 14: Pipeline of equipping the model with 3D spatial cues.

D.2 Chain-of-Thought Prompting

To address issues such as Prompt Alignment and Latent Logic Inference errors, we explored Chain-of-Thought (CoT) prompting, guiding models to reason step by step. The model is provided with explicit prompts for each step:

- *Step1: Understand and Analyze. Interpret the problem input, including auxiliary visual inputs, preset conditions, or requirements, and identify the key information to extract from the video.*
- *Step2: Locate and Gather Evidence. Find the relevant information in the video and collect sufficient evidence, including implicit clues not directly mentioned in the input.*
- *Step3: Reason and Solve. Combine the prompt with the extracted video information and perform step-by-step reasoning to answer the question.*

As shown in Fig. 15, simply encouraging the model to "think step by step" does not consistently improve performance. This aligns with previous findings (Yang et al., 2024; 2025). The underlying issue is not that the model forgets to perform certain steps, but rather that it struggles to handle inherently difficult aspects of the task, highlighting that the limitation lies in the model's intrinsic reasoning ability.

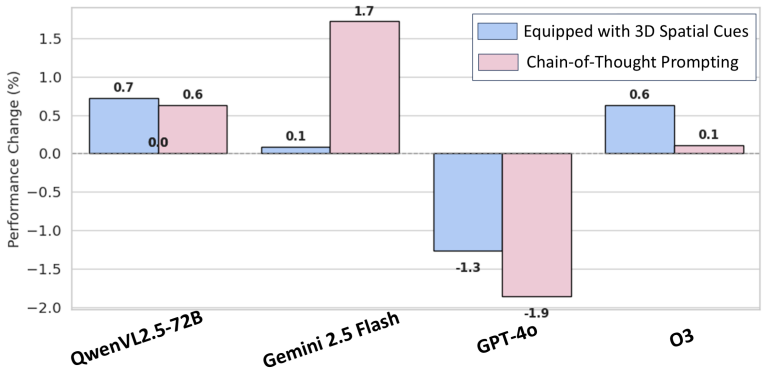

Figure 15: Effect of different methods on model performance.

D.3 Suggestions for Model Improvement.

We explored two training-free strategies to enhance models' spatial reasoning abilities, including prompt-based cue emphasis and explicit instruction to utilize the provided 3D spatial clues. Neither approach led to noticeable performance improvements, suggesting that the observed limitations are not due to suboptimal prompting or inference-time heuristics, but rather stem from more fundamental deficiencies in current MLLMs. Specifically, existing models lack architectural mechanisms that can naturally encode and reason over 3D geometric structures. As a result, when provided without additional training, spatial cues are treated as peripheral information rather than being integrated into the core reasoning process. This limitation cannot be effectively addressed through training-free methods alone.

Therefore, improving model performance on our benchmark requires progress along two complementary directions. On the architectural side, models should incorporate inductive biases or explicit modules for 3D geometry, spatial continuity, and viewpoint transformation, enabling more accurate representation and manipulation of 3D relationships. On the data side, large-scale, diverse training data with rich spatial supervision are necessary to align visual representations with geometric reasoning and to encourage consistent utilization of spatial cues, rather than relying on the limited and fixed-form training data, which makes it difficult for models to generalize across a wide range of benchmarks as discussed in Sec. 4.3. We believe that advances along both axes are essential for closing the gap between current MLLMs and truly robust spatial intelligence.

E Social Impact

E.1 Positive Social Impact

MMSI-Video-Bench contributes to advancing spatial intelligence in MLLMs by providing a comprehensive and rigorously constructed benchmark that systematically reveals their current limitations in spatial reasoning over video. By exposing a substantial human–AI gap and identifying consistent failure modes—such as geometric reasoning, motion grounding, long-horizon prediction, and cross-video correspondence—it establishes a clear and standardized testbed for measuring progress. We expect this benchmark to facilitate the development of more reliable and generalizable spatial intelligence, with downstream benefits for applications such as robotics, embodied agents, autonomous systems, and immersive environments, where accurate spatial understanding is essential.

E.2 Negative Social Impact

As a question-answering benchmark, MMSI-Video-Bench is primarily suited for evaluating models with compatible interfaces and may not fully capture the capabilities of non-QA-based systems. Misuse or overgeneralization of its results beyond this scope could lead to misleading conclusions about model performance.

F License Information for the Public Datasets Used

As mentioned in the main paper, our real-world scene data is sourced from Roomtour3D (Han et al., 2025), ScanNet (Dai et al., 2017), ScanNet++ (Yeshwanth et al., 2023), 3RScan (Wald et al., 2019), ARKitScenes (Baruch et al., 2021), RealEstate10K (Zhou et al., 2018), DL3DV (Ling et al., 2024), Waymo (Sun et al., 2020), nuScenes (Caesar et al., 2020), OVIS (Qi et al., 2022), TrackingNet (Müller et al., 2018), LaSOT (Fan et al., 2020), UAV123 (Mueller et al., 2016), Ego4D (Grauman et al., 2022), EPIC-KITCHENS (Damen et al., 2018), EgoExoLearn (Huang et al., 2024), MultiSports (Li et al., 2021), Charades (Sigurdsson et al., 2016), LEMMA (Jia et al., 2020), TF2023 (Zhao et al., 2024), CVMHT (Han et al., 2020), AVA (Gu et al., 2018), DROID (Khazatsky et al., 2024), RH20T (Fang et al., 2024), and DTU (Jensen et al., 2014). To access and use these datasets, users are required to comply with the original licenses and terms of use specified on their respective official websites, as summarized in Tab. 11. We utilize these datasets as scene data and leverage the video information they provide. Throughout their use, all licensing requirements and terms of use are strictly adhered to.

Table 11: Public datasets used in our benchmark with licenses and access URLs.

Dataset	License	URL
Roomtour3D	cc-by-sa-4.0	https://huggingface.co/datasets/roomtour3d/room_tour_video_3fps
ScanNet	Custom (sca, a)	http://www.scan-net.org/
ScanNet++	Custom (sca, b)	https://scannetpp.mlsg.cit.tum.de/scannetpp/
3RScan	Custom(research-only)	https://waldjohannau.github.io/RIO/
ARKitScenes	Apple License	https://machinelearning.apple.com/research/arkitscenes
RealEstate10K	cc-by-4.0	https://huggingface.co/datasets/RE10K/RealEstate10K
DL3DV	Custom (dl3)	https://huggingface.co/DL3DV
Waymo	Non-Commercial Use	https://waymo.com/open/
nuScenes	cc-by-sa-4.0	https://www.nuscenes.org/
OVIS	cc-by-4.0	https://songbai.site/ovis/
TrackingNet	gpl-3.0	https://huggingface.co/datasets/SilvioGiancola/TrackingNet
Ego4D	Custom (ego)	https://ego4d-data.org/
EPIC-KITCHENS	cc-by-nc-4.0	https://epic-kitchens.github.io/2026
MultiSports	cc-by-nc-4.0	https://huggingface.co/datasets/MCG-NJU/SportsAction
Charades	Non-Commercial Use	https://huggingface.co/datasets/HuggingFaceM4/charades
LEMMA	Custom(research-only)	https://sites.google.com/view/lemma-activity
TF2023	Custom(research-only)	https://github.com/ziweizhao1993/PEN
CVMHT	Custom(research-only)	https://github.com/RuizeHan/CVMHT
AVA	cc-by-4.0	https://research.google.com/ava/
DROID	apache-2.0	https://droid-dataset.github.io/
RH20T	cc-by-sa-4.0	https://rh20t.github.io/#download
DTU	Custom(research-only)	https://roboimagedata.compute.dtu.dk/

G Additional Error Analysis

G.1 Error Category Definitions

Effective video understanding requires a sequence of reasoning steps: first perceiving fine-grained details, then linking entities across frames, followed by modeling spatial relations, and finally correctly aligning prompts to answer questions, with some cases demanding deeper reasoning over implicit cues. Based on this structured process, we categorize all model errors in our bench into non-overlapping, comprehensive types (Fig. 4):

Detailed Grounding Error. Failures in fine-grained perception, including missing or confusing objects, overlooking subtle temporal changes, or misidentifying events at specific timestamps. This error mainly reflects deficiencies in surface-level visual grounding.

ID Mapping Error. Failures in maintaining consistent identity tracking across frames, often caused by occlusion, rapid motion, or visually similar distractors, leading the model to confuse or mismatch entities over time.

Geometric Reasoning Error. Mistakes in inferring spatial relations (relative positions or distance, e.g., front/behind, near/far), revealing the model's inability to establish coherent spatial associations across frames.

Prompt Alignment Error. Misunderstandings in interpreting the prompt or integrating it with visual evidence. These occur when the prompt introduces new conditions, reference images, or auxiliary visual inputs that the model fails to correctly incorporate, even if its understanding of the video information itself is accurate.

Latent Logical Inference Error. Failures in reasoning that require integrating implicit cues or commonsense knowledge. Some questions in MMSI-Video-Bench demand inference based on subtle contextual clues, such as choosing an appropriate reference object to estimate height/ distance/ speed or correlating information across different viewpoints, or predicting motion trajectories using basic physical intuition. Models fail to detect or leverage these implicit cues.