MMSI-Video-Bench: A Holistic Benchmark for Video-Based Spatial Intelligence

Jingli Lin[1,2*] **Runsen Xu**[1,3*†] **Shaohao Zhu**[1,4] **Sihan Yang**[1] **Peizhou Cao**[1,5]
Yunlong Ran[1,4] **Miao Hu**[6] **Chenming Zhu**[1,7] **Yiman Xie**[1,4] **Yilin Long**[1,8] **Wenbo Hu**[1,9]
Dahua Lin[1,3] **Tai Wang**[1✉] **Jiangmiao Pang**[1✉]

[1]Shanghai AI Laboratory [2]Shanghai Jiaotong University [3]The Chinese University of Hong Kong [4]Zhejiang University [5]Beihang University [6]Xi'an Jiaotong University [7]University of Hong Kong [8]Fudan University [9]University of California, Los Angeles

[*] **Equal Contribution** [†] **Project Lead**

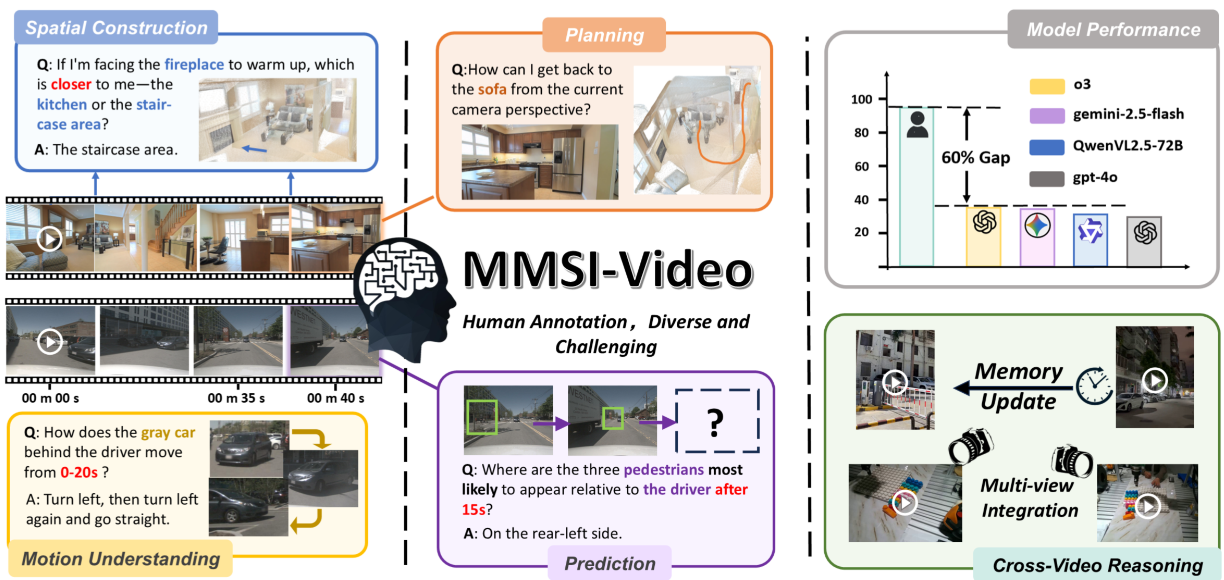

Figure 1. **MMSI-Video-Bench** is a diverse, human-annotated, and challenging benchmark, designed to evaluate models' video-based spatial intelligence, including their ability to perceive, understand, reason, and make decisions over spatio-temporal information in videos. The bar chart in the top-right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

Abstract

*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 1,107 questions grounded in 1,279 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 22 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%. 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 benchmark due to its reasoning-intensive nature. We expect our benchmark to establish a solid testbed for advancing video-based spatial intelligence.*

1. Introduction

For decades, humans have aspired to build an embodied general-purpose AI assistant, akin to JARVIS in *Iron Man*. As MLLMs [1, 5, 45, 51, 54] 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, that is, 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 [4, 6, 24] or multiple images [14, 22, 47] rather than videos, leaving an input gap to the practical setting. Recent video-based benchmarks [28, 29, 46, 50] 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 [47]; 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 Figure 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 [47]. 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 1,107 questions grounded in 1,279 video clips, grouped into five task categories and 13 subtypes.

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, OpenAI's o3 [33], 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.

To provide diagnostic signals for future research, we perform per-category error analysis. For Spatial Construction, failures are dominated by geometric reasoning errors; for Motion, by fine-grained grounding failures on fast, subtle, or long-duration motion; for Planning and Prediction, by prompt–evidence misalignment where models ignore video cues; and for Cross-Video Reasoning, by difficulty grounding and matching correspondences across videos. We further observe that, due to the reasoning-intensive nature of our benchmark, the commonly used frame-sampling strategy AKS [37] does not transfer directly to MMSI-Video-Bench, leading to additional performance degradation. Overall, MLLMs still have substantial room for improvement in spatial intelligence, and MMSI-Video-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 [44], and ActivityNet-QA [49] mainly assess global visual comprehension, with limited attention to temporal understanding. Later works like NeXT-QA [43] and MVBench [25] emphasize temporal dynamics. Subsequently, benchmarks such as LongVideoBench [42] and Video-MME [13] 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 [28, 29, 46, 50], online inference under continuous observation [27, 29], and cross-video reasoning [18]. 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 [6], SpatialVLM [4], and CVBench [38] focus on single-image spatial reasoning, emphasizing depth and distance. Later, video-based benchmarks, VSI-Bench [46], SPAR-Bench [50], and OST-Bench [29], extend this to indoor scenes with object–object and object–camera rela-

Benchmark	Source Diversity	Modality	Annotation Method	Task	Samples Num	Human-AI Gap
SpatialRGPT [6]	-	Single-Image	Auto	local-SU.	1,406	<42
SpatialVLM [4]	-	Single-Image	Auto&Human	local-SU.	546	<30
CVBench [38]	3	Single-Image	Auto	local-SU.	2,638	-
All-Angles-Bench [7]	2	Multi-Image	Human	SU.	2100	21.2
MMSI-Bench [47]	8	Multi-Image	Human	local-SU. & short-MU.	1,000	55.3
VSI-Bench [46]	3	Video	Auto	SU. & Plan.	5,000	33
OST-Bench [29]	3	Video	Auto	SU. & CAM.-MU.	10,000	29.3
SPAR-Bench [50]	3	Video	Auto	SU. & CAM.-MU.	7207	27.8
STI-Bench [28]	3	Video	Auto	SU. & INST./CAM.-MU.	2,064	-
EgoExoBench [18]	6	Multi-Video	Auto&Human	CV.	7,330	41.9
MMSI-Video (ours)	26	Video/Multi-Video	Human	SU. & MU. & Plan./Pred. & CV.	1,107	59.2

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. The Human–AI Gap indicates the performance difference as a percentage.

tions, yet remain constrained by limited scene diversity and static spatial contexts. More recent benchmarks, including MMSI-Bench [47] and STI-Bench [28], incorporate dynamic environments and cover a wider range of indoor–outdoor scenarios. MMSI-Bench serves as a demanding benchmark 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.

3. MMSI-Video-Bench

In this section, we introduce the formulation of our MMSI-Video-Bench and the methodology for its construction.

3.1. Overview

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. The *Planning* and *Prediction* categories focus on higher-level reasoning over sparse video information:(1) Planning tasks require the model to devise actions toward a specific goal based on visual cues; (2) 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 often 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–temporal 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 the supplementary material, we provide an overview table that describes the details of each subtype.

3.2. Benchmark Construction

Data Collection & Preprocessing. To ensure diversity in our benchmark, we curated videos spanning a broad spectrum of real-world scenarios. In terms of capture types, the collection covers tabletop recordings, single-room indoor scenes, multi-room and multi-floor indoor environments, outdoor building and street views, natural outdoor settings, sports activities, and movie footage. Our data sources include 25 publicly available datasets [2, 3, 8–10, 16, 20,

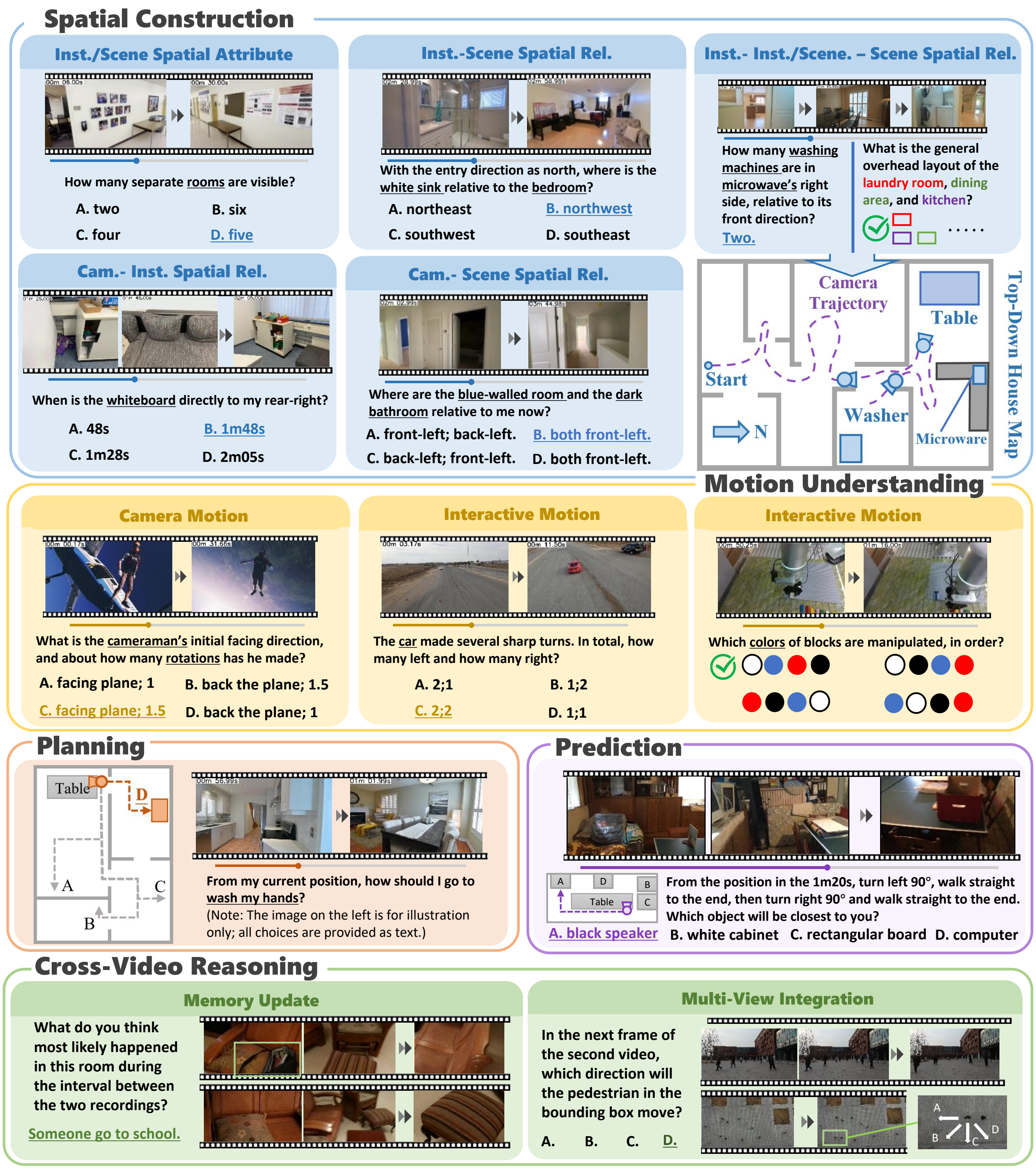

Figure 2. Illustrative examples of different subtypes. Rel., Inst., and Cam. stand for Relationship, Instance, and Camera.

31, 32, 39, 48], such as ScanNet [8], RealEstate10k [53], DL3DV [30], Ego4D [15], DROID [23], Waymo [36] and MultiSports [26] (the remaining datasets are listed in the supplementary material), amounting to approximately 20k video clips. In addition, we manually recorded and carefully selected 140 supplemental 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 (formatted as "xx min yy s") to facilitate precise temporal referencing during question design.

Human Annotation. All annotations were conducted by a team of eleven trained researchers specializing in 3D vi-

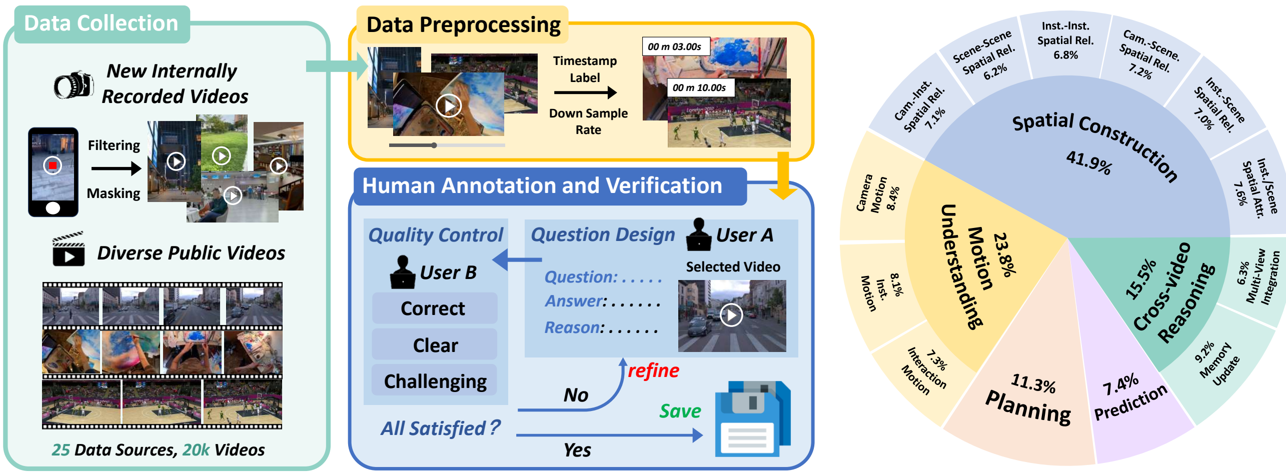

Figure 3. (Left) The pipeline of benchmark sample construction. (Right) The distribution of question types in MMSI-Video-Bench.

sion. 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.

Statistics. After the construction of benchmark samples, the final MMSI-Video-Bench contains 1,107 questions grounded in 1,279 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. Additional benchmark statistics are provided in the supplementary material.

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 o3/o4-mini/gpt-4o, gemini-2.5-flash, claude-4.5, 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. In this setting, each model receives the complete set of frames used during annotation, ensuring 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.

Besides general-purpose models, we also evaluate several models that are finetuned on spatial reasoning data or equipped with latent spatial representations. Their results and analyses are provided in the supplementary material.

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, o3, lags behind humans (94.22) by nearly 60%. This indicates that current models are unable to handle these challenging tasks, highlighting the inherent

Models	Spatial Construction						Motion Understanding			Cross-Video		Plan.	Pred.	Avg.
	Attr.	Inst.-Inst.	Inst.–Scen.	Scen.–Scen.	Cam.-Inst.	Cam.–Scen.	Cam.	Inst.	Inter.	MU.	MV.	-	-	
(Sufficient-Coverage) Proprietary														
gpt-4o	31.3	34.2	20.8	30.4	20.2	33.8	33.3	26.7	**29.6**	32.4	**37.1**	34.4	20.7	29.8
o3	37.4	**38.2**	**45.5**	**43.5**	32.9	38.8	32.3	28.9	**29.6**	31.4	35.7	32.8	**34.1**	**35.0**
o4-mini	31.3	28.9	39.0	37.7	34.2	32.5	**36.6**	18.9	25.9	**37.2**	31.4	**37.6**	30.5	32.6
gemini-2.5-flash	**43.4**	27.6	41.6	**43.5**	**40.5**	41.2	32.3	27.8	27.2	33.3	30.0	32.8	28.1	34.3
-Thinking	42.2	26.3	42.9	31.9	25.3	**51.2**	34.4	**32.2**	25.9	35.3	34.3	36.0	26.8	34.3
(Sufficient-Coverage) Open-source														
InternVL2.5-8B	28.9	28.9	26.0	27.5	13.9	31.2	23.7	24.4	29.6	31.4	28.6	17.6	29.3	25.9
InternVL3-8B	24.1	32.9	27.3	30.4	19.0	25.0	**34.4**	32.2	27.2	22.6	31.4	28.8	23.2	27.6
InternVideo2.5-8B	27.7	26.3	26.0	30.4	19.0	**32.5**	24.7	32.2	27.2	29.4	30.0	24.8	28.1	27.5
QwenVL2.5-7B	28.9	**34.2**	24.7	34.8	20.2	28.8	33.3	23.3	27.2	25.5	30.0	29.6	30.5	28.5
QwenVL2.5-32B	32.5	23.7	27.3	**36.2**	22.8	25.0	31.2	**34.4**	33.3	33.3	35.7	**32.8**	21.9	30.2
QwenVL2.5-72B	**36.1**	25.0	28.6	23.2	**31.6**	31.2	31.2	**34.4**	30.9	31.4	**38.6**	27.2	**35.4**	**31.1**
QwenVL3-8B	26.5	28.9	27.3	29.0	29.1	30.0	25.8	27.8	**35.8**	**36.3**	28.6	28.0	26.8	29.3
QwenVL3-30B	25.3	30.3	22.1	24.6	21.5	31.2	26.9	25.6	29.6	32.4	27.1	31.2	26.8	27.6
-Thinking	21.7	27.6	**35.1**	27.5	22.8	23.8	30.1	28.9	32.1	35.3	24.3	29.6	26.8	28.4
(Uniform-50) Proprietary														
claude-haiku-4-5	30.1	32.9	23.4	30.4	26.6	36.2	37.6	34.4	28.4	37.2	31.4	26.4	**34.1**	31.5
gpt-4o	33.7	34.2	28.6	29.0	29.1	35.0	30.1	**35.6**	**35.8**	25.5	30.0	33.6	28.1	31.4
gpt-5	**42.2**	**38.2**	35.1	39.1	38.0	38.8	**40.9**	33.3	30.9	36.3	**40.0**	32.8	25.6	36.0
o3	39.8	34.2	40.3	**44.9**	38.0	33.8	36.6	31.1	32.1	**38.2**	31.4	**39.2**	30.5	**36.2**
o4-mini	32.5	36.8	40.3	33.3	**43.0**	40.0	36.6	33.3	28.4	34.3	32.9	32.0	31.7	34.9
gemini-2.5-flash	**42.2**	34.2	**41.6**	33.3	35.4	42.5	36.6	26.7	24.7	30.4	25.7	35.2	29.3	33.7
-Thinking	39.8	34.2	39.0	33.3	25.3	**47.5**	36.6	28.9	25.9	30.4	32.9	36.0	30.5	33.9
(Uniform-50) Open-source														
InternVL2.5-8B	24.1	32.9	32.5	24.6	16.5	27.5	20.4	34.4	35.8	30.4	31.4	28.8	23.2	27.9
InternVL2.5-38B	31.3	32.9	19.5	23.2	16.5	**37.5**	28.0	23.3	34.6	**38.2**	31.4	32.0	**35.4**	29.8
InternVL2.5-78B	30.1	34.2	28.6	30.4	22.8	**37.5**	31.2	30.0	29.6	34.3	28.6	29.6	28.1	30.4
InternVL3-8B	30.1	28.9	**35.1**	18.8	16.5	23.8	26.9	25.6	33.3	30.4	31.4	30.4	31.7	28.1
InternVL3-38B	28.9	35.5	32.5	31.9	21.5	30.0	**36.6**	33.3	24.7	30.4	25.7	25.6	24.4	29.3
InternVL3-78B	**37.4**	31.6	33.8	24.6	26.6	26.2	28.0	34.4	**40.7**	37.2	**37.1**	**35.2**	34.1	**33.1**
InternVideo2.5-8B	33.7	25.0	27.3	26.1	19.0	23.8	29.0	**35.6**	18.5	31.4	30.0	24.8	32.9	27.6
LLaVA-Video-7B	32.5	27.6	26.0	**37.7**	29.1	28.8	34.4	27.8	25.9	35.3	22.9	25.6	30.5	29.5
LLaVA-Video-72B	30.1	**38.2**	32.5	29.0	13.9	25.0	28.0	33.3	33.3	35.3	34.3	24.0	31.7	29.7
QwenVL2.5-7B	28.9	23.7	29.9	26.1	25.3	31.2	28.0	26.7	24.7	27.4	21.4	28.0	34.1	27.5
QwenVL2.5-32B	26.5	23.7	31.2	24.6	16.5	25.0	30.1	27.8	34.6	31.4	28.6	34.4	29.3	28.4
QwenVL2.5-72B	31.3	31.6	28.6	23.2	25.3	32.5	33.3	**35.6**	**40.7**	32.4	**37.1**	24.8	28.1	31.0
QwenVL3-8B	36.1	34.2	26.0	24.6	19.0	23.8	29.0	21.1	29.6	28.4	32.9	29.6	28.1	27.9
QwenVL3-30B	27.7	32.9	29.9	26.1	16.5	32.5	26.9	28.9	29.6	27.4	30.0	26.4	28.1	27.8
-Thinking	28.9	30.3	29.9	34.8	**31.6**	28.8	28.0	26.7	**40.7**	32.4	24.3	27.2	19.5	29.4
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	90.1	94.4	89.0	95.1	92.9	94.2

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

difficulty of MMSI-Video-Bench. This observation motivates a deeper investigation into the underlying reasons for the models' subpar performance on this benchmark.

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 [29, 46, 47]. 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 the Uniform-50 setting, 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 [40] 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 presented in the frame sampling study section.

Comparison across Models. We observe that proprietary models consistently outperform open-source ones. Among open-source models, the best-performing ones, InternVL3-78B (Uniform-50) and QwenVL2.5-72B (Sufficient-Coverage), still exhibit a noticeable gap com-

Method	GPT-4o	QwenVL2.5-72B	InternVL3-78B
Uniform-50	31.4	31.0	33.1
AKS-50	29.7 (-1.7)	31.1 (+0.1)	32.6 (-0.5)

Table 3. Comparison of model performance under uniform sampling and AKS sampling strategies.

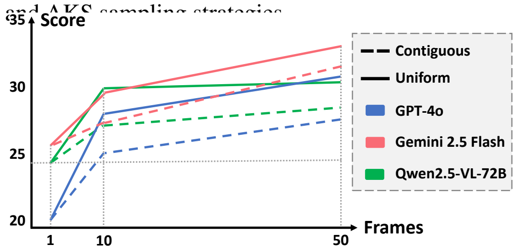

Figure 4. Model performance under different frame sampling methods. Dashed lines indicate contiguous sampling, while solid lines indicate uniform sampling.

pared 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 with their thinking mode versions.

5. 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.4. 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 outperforms 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.

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 [37], which selects frames based on image–text semantic representations, we sampled 50 frames per video and evaluated model performance. Results are summarized in Table 3. Although AKS achieves substantial gains on benchmarks such as LongVideoBench [42] and Video-MME [13], 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.

6. Error Analysis

6.1. Error Categorization

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.5):

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. The model fails to detect or leverage these implicit cues.

6.2. Error Statistics

We select four representative models (gpt-4o, gemini-2.5-flash, o3, and QwenVL2.5-72B) and conducted an error

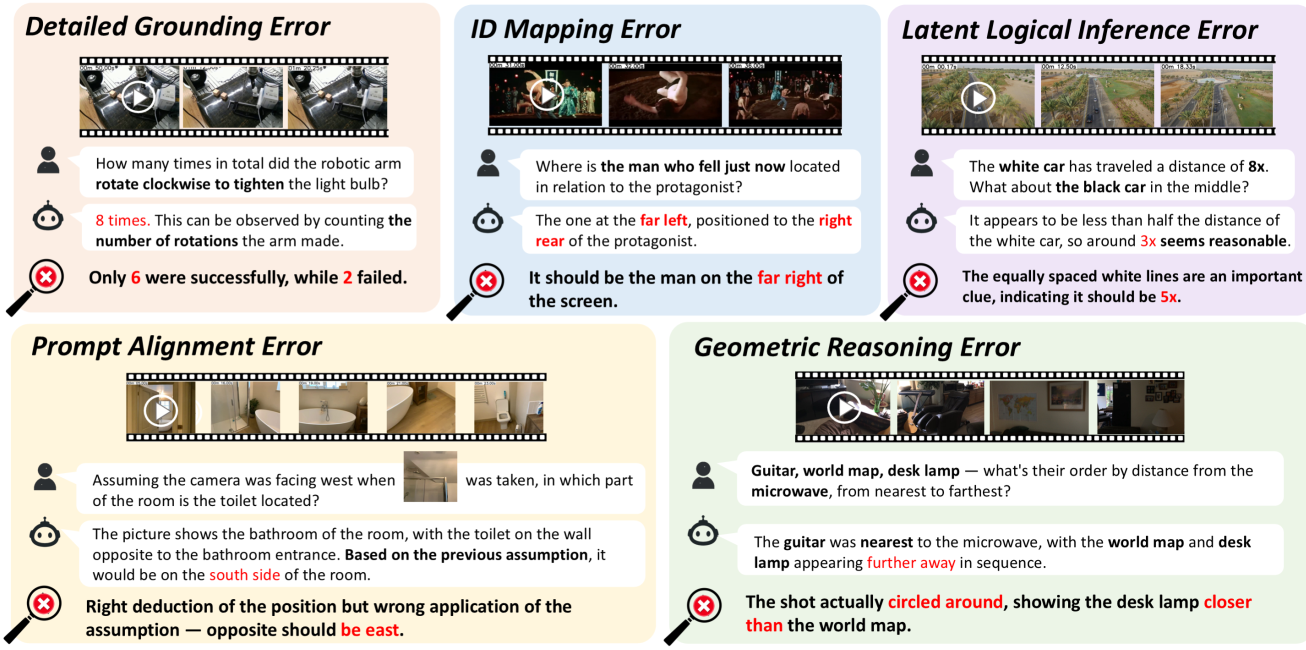

Figure 5. Illustration of five representative error types identified in MMSI-Video-Bench, along with examples of model responses and corresponding error analyses.

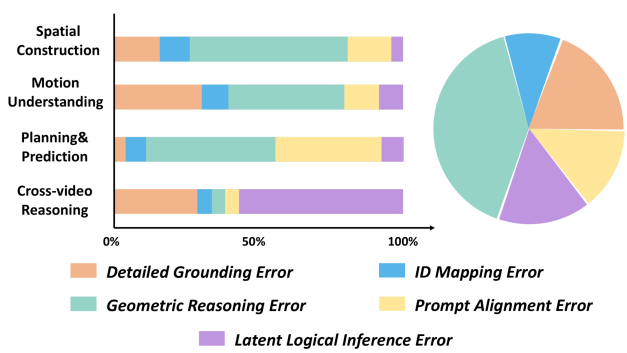

Figure 6. Distribution of the five error types. (Left) Error distribution across different question categories. (Right) Overall proportion of each error type.

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. 6, 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:

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[29, 46, 47], indicating that current models still struggle with inferring even simple geometric relations.

Distinct error distributions reveal task-specific capability bottlenecks. Beyond Spatial Construction, we observe the following patterns across other task categories:

- **In Motion Understanding tasks**, detailed grounding remains a major limitation: models often fail to comprehensively detect or interpret motion patterns, especially when confronted with fast movements, subtle actions, or long-duration motions.
- **In Planning and Prediction tasks**, Prompt Alignment Error is a significant source of issues: models may accurately perceive the spatiotemporal context but still fail to connect high-level goals, assumptions, or contextual conditions with the video evidence.
- **In Cross-Video Reasoning tasks**, Latent Logical Inference Errors are most prominent, followed by Detailed Grounding Errors. These tasks typically require identifying correspondences across multiple videos (i.e., using matching instances across videos from different time points or viewpoints to establish spatio-temporal correspondences between the videos). We find that models frequently either fail to locate the same instance in both videos simultaneously or neglect to utilize them effectively for reasoning.

Through this error analysis, we gain a deeper understanding of the specific failure modes associated with each category in MMSI-Video-Bench, offering valuable insights into which model capabilities require improvement and which weaknesses future iterations should target. Based on these findings, we provide discussions and preliminary explorations on potential ways to improve model performance in the supplementary material.

7. Conclusion

We present MMSI-Video-Bench, a diverse, human-annotated, holistic video-based spatial intelligence benchmark that comprehensively evaluates models' perception, understanding, reasoning, and decision-making over spatiotemporal information. Our evaluation reveals a substantial gap between model and human performance, with models struggling across all task categories beyond spatial construction. Error analysis reveals task-specific failure patterns, showing where current models struggle and providing concrete guidance for improving different aspects of spatial intelligence; and the frame sampling study underscores the benchmark's challenge and emphasizes the need for a more effective sampling strategy. Overall, MMSI-Video-Bench provides a rigorous and holistic testbed for evaluating models' spatial intelligence, while our analysis offers actionable insights and directions for future improvements.

References

[1] 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. 2

[2] 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. 3, 1

[3] 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. 3, 1

[4] 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. 2, 3

[5] 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. 2

[6] 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. 2, 3

[7] Shengbang Tong Ta-Ying Cheng Rouyu Wang Tianzhe Chu Yuexiang Zhai Yubei Chen Shenghua Gao Chun-Hsiao Yeh, Chenyu Wang and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in mllms. *arXiv preprint arXiv:2504.15280*, 2025. 3

[8] 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*, pages 5828–5839, 2017. 3, 4, 1

[9] 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. 1

[10] 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. 3, 1

[11] 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. 5

[12] 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)*, pages 653–660. IEEE, 2024. 1

[13] 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*, 2024. 2, 7

[14] 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, 2024. 2

[15] 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 Fue-

gen, 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. 4, 1

[16] 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. 3, 1

[17] 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. 1

[18] 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. 2, 3

[19] 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. 1

[20] 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*, pages 406–413. IEEE, 2014. 3, 1

[21] 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. 1

[22] 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. 2, 5

[23] 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. 4, 1

[24] 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. 2

[25] 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. 2

[26] 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)*, pages 13536–13545, 2021. 4, 1

[27] 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?, 2025. 2

[28] Yun Li, Yiming Zhang, Tao Lin, XiangRui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. Sti-bench: Are mllms ready for precise spatial-temporal world understanding? *arXiv preprint arXiv:2503.23765*, 2025. 2, 3

[29] 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. 2, 3, 6, 8, 5

[30] 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*, pages 22160–22169, 2024. 4, 1

[31] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In *Computer Vision – ECCV 2016*, pages 445–461, Cham, 2016. Springer International Publishing. 4, 1

[32] 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. 4, 1

[33] OpenAI, :, Ahmed El-Kishky, Alexander Wei, Andre Saraiva, Borys Minaiev, Daniel Selsam, David Dohan, Francis Song, Hunter Lightman, Ignasi Clavera, Jakub Pachocki,

Jerry Tworek, Lorenz Kuhn, Lukasz Kaiser, Mark Chen, Max Schwarzer, Mostafa Rohaninejad, Nat McAleese, o3 contributors, Oleg Mürk, Rhythm Garg, Rui Shu, Szymon Sidor, Vineet Kosaraju, and Wenda Zhou. Competitive programming with large reasoning models, 2025. 2

[34] 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. 1

[35] 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. 1

[36] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott 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)*, 2020. 4, 1

[37] Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding, 2025. 2, 7

[38] CVBench Team. Cvbench: A benchmark for cross-video multimodal reasoning, 2025. 2, 3

[39] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance relocalization in changing indoor environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7658–7667, 2019. 4, 1

[40] 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, 2025. 6

[41] 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. 5

[42] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024. 2, 7

[43] 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)*, pages 9777–9786, 2021. 2

[44] 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*, pages 1645–1653, 2017. 2

[45] 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. 2

[46] 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. 2, 3, 6, 8

[47] 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. 2, 3, 6, 8, 5

[48] 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*, pages 12–22, 2023. 4, 1

[49] 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*, pages 9127–9134, 2019. 2

[50] 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*, 2025. 2, 3

[51] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data, 2025. 2

[52] 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)*, pages 16477–16487, 2024. 1

[53] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images, 2018. 4, 1

[54] 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. 2

MMSI-Video-Bench: A Holistic Benchmark for Video-Based Spatial Intelligence

Supplementary Material

8. Benchmark Details

8.1. 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.13.

8.2. Data Collection & Preprocessing Details

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.

Dataset	Type	FPS	Duration(Sec.)
Roomtour3d [16]	Indoor Scan.	1.0	466.86
ScanNet [8]	Indoor Scan.	1.0	39.92
ScanNet++ [48]	Indoor Scan.	2.0	136.35
3RScan [39]	Indoor Scan.	1.0	60.94
ARKitScenes [2]	Indoor Scan.	1.0	77.28
RealEstate10k [53]	Indoor Scan.	0.66	214.73
DL3DV [30]	Indoor&Outdoor	1.0	39.81
Waymo [36]	Outdoor Env.	5.0	15.92
NuScenes [3]	Outdoor Env.	4.0	26.79
OVIS [34]	Outdoor Env.	5.0	13.74
TrackingNet [32]	Outdoor Env.	4.0	21.33
LaSOT [10]	Outdoor Env.	5.89	32.40
UAV123 [31]	Outdoor Env.	5.89	24.32
Ego4D [15]	Ego.-Int.	2.0/8.33	262.49
EPIC-KITCHENS [9]	Ego.-Int.	2.0/8.33	91.51
EgoExoLearn [19]	Ego.-Int.	4.0	565.81
MultiSports [26]	Exo.-HA.	2.0/8.33	20.51
charades [35]	Exo.-HA.	4.0	27.91
LEMMA [21]	Exo.-HA.	12.50	22.84
TF2023 [52]	Exo.-HA.	2.0	492.46
CVMHT [17]	Exo.-HA.	4.0	37.25
AVA [17]	Exo.-HA.	1.0	900.27
DROID [23]	Others	4.0	86.75
RH20T [12]	Others	4.0	84.32
DTU [20]	Others	2.0	24.00
RealWorld	Indoor&Outdoor	2.0	46.43

Table 4. Statistics of all source video datasets, including their capture types, average durations, and standardized FPS after preprocessing.Scan./Env. denotes scanning, environment.; Exo.-Int. denotes egocentric interactions and Exo.-HA. denotes exocentric human activities.

8.3. Human Annotation UI

As shown in Fig.12, 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

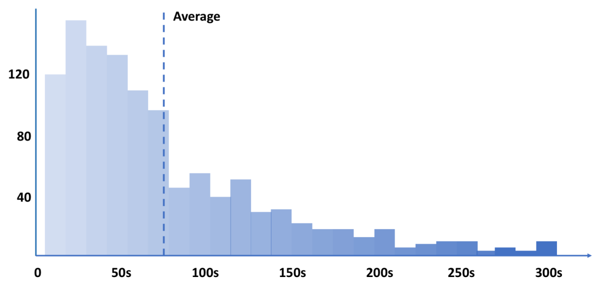

Figure 7. Duration distribution of all video samples in MMSI-Video-Bench (in seconds).

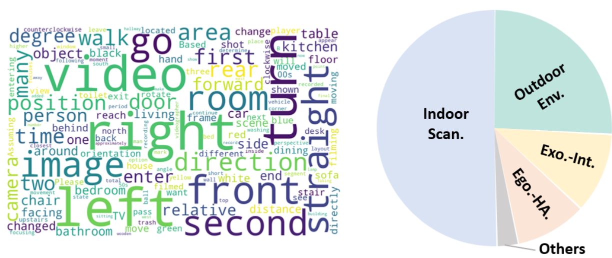

Figure 8. (Left) Word cloud of MMSI-Video-Bench. (Right) Distribution of video source categories across all samples in MMSI-Video-Bench.

samples, the validator is required to provide reasons and suggestions for revision.

8.4. More Statistics

As shown in Fig.7, we present the distribution of video durations across all samples in MMSI-Video-Bench. Fig.8 further illustrates a word cloud of the benchmark annotations, together with the distribution of video source capture types.

9. Experiment Details

In our evaluation, all models are provided with the same input template. As illustrated in Fig.14, 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.

10. More Analysis

10.1. More Findings of the Main Results

Prediction is the most challenging main category, and Camera–Instance Spatial Relation is the most challenging subtype. Based on the results presented in the main paper, 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. Based on the results presented in the main paper, we computed the average scores of each model across the main categories, as shown in Tab.5. Overall, o3 achieves the highest performance, while gemini-2.5-flash demonstrates the strongest capabilities in spatial construction and reasoning tasks. On the other hand, gpt-5 excels at motion understanding and reasoning across multiple video segments. As a model with high reasoning and decision-making capacity, o3 also leads in Prediction and Planning tasks compared to other models. 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 claude-haiku-4-5 and gemini-2.5-flash are particularly effective at solving questions that most other models struggle with.

10.2. 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 examined not only through the categorization presented in the main paper but also 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

Methods	Avg	SC.	MU.	Plan.&Pred.	CV.	Hard	Medium	Easy
(Sufficient-Coverage) Proprietary								
gpt-4o	29.8	28.4	29.9	29.0	34.3	15.9	27.0	44.7
o3	**35.0**	39.2	30.3	33.3	33.1	15.3	30.2	**57.1**
o4-mini	32.6	33.8	27.3	**34.8**	**34.9**	13.2	29.5	52.6
gemini-2.5-flash	34.3	**39.7**	29.2	30.9	32.0	17.1	**31.8**	51.8
-Thinking	34.3	36.9	**31.1**	32.4	**34.9**	**17.4**	31.0	52.4
(Sufficient-Coverage) Open-source								
InternVL2.5-8B	25.9	26.1	25.8	22.2	30.2	9.8	22.8	43.2
InternVL3-8B	27.6	26.3	31.4	26.6	26.2	**13.8**	23.2	44.0
InternVideo2.5-8B	27.5	26.9	28.0	26.1	29.6	8.0	25.0	46.8
QwenVL2.5-7B	28.5	28.4	28.0	29.9	27.3	12.2	**27.5**	43.4
QwenVL2.5-32B	30.2	27.8	**33.0**	28.5	**34.3**	13.2	23.2	52.1
QwenVL2.5-72B	**31.1**	**29.5**	32.2	**30.4**	**34.3**	11.3	25.5	**54.0**
QwenVL3-8B	29.3	28.4	29.6	27.5	33.1	8.9	25.5	50.8
QwenVL3-30B	27.6	25.9	27.3	29.5	30.2	7.0	23.2	49.7
-Thinking	28.4	26.3	30.3	28.5	30.8	10.1	24.2	48.4
(Uniform-50) Proprietary								
claude-haiku-4-5	31.5	30.0	33.7	29.5	34.9	**16.8**	27.2	48.7
gpt-4o	31.4	31.7	33.7	31.4	27.3	6.4	29.5	55.0
gpt-5	36.0	**38.6**	**35.2**	29.9	**37.8**	14.7	**36.2**	54.2
o3	**36.2**	38.4	33.3	**35.8**	35.5	11.6	34.0	59.7
o4-mini	34.9	37.7	33.0	31.9	33.7	10.4	31.8	59.2
gemini-2.5-flash	33.7	38.4	29.6	32.9	28.5	8.3	29.5	**60.0**
-Thinking	33.9	36.6	30.7	33.8	31.4	7.7	33.8	56.6
(Uniform-50) Open-source								
InternVL2.5-8B	27.9	26.3	29.9	26.6	30.8	6.7	22.8	51.6
InternVL2.5-38B	29.8	26.9	28.4	33.3	35.5	9.5	23.2	54.2
InternVL2.5-78B	30.4	**30.6**	30.3	29.0	32.0	12.5	24.0	52.6
InternVL3-8B	28.1	25.6	28.4	30.9	30.8	7.3	25.8	48.4
InternVL3-38B	29.3	30.0	31.8	25.1	28.5	8.9	25.5	50.8
InternVL3-78B	**33.1**	30.2	34.1	**34.8**	**37.2**	**13.5**	**27.8**	55.5
InternVideo2.5-8B	27.6	25.9	28.0	28.0	30.8	6.1	23.2	50.5
LLaVA-Video-7B	29.5	30.2	29.6	27.5	30.2	8.3	23.8	54.0
LLaVA-Video-72B	29.7	28.0	31.4	27.1	34.9	12.2	23.5	51.3
QwenVL2.5-7B	27.5	27.6	26.5	30.4	25.0	6.4	23.2	50.0
QwenVL2.5-32B	28.4	24.6	30.7	32.4	30.2	7.7	21.2	53.7
QwenVL2.5-72B	31.0	28.9	**36.4**	26.1	34.3	7.3	23.5	**59.2**
QwenVL3-8B	27.9	27.4	26.5	29.0	30.2	4.9	21.8	54.2
QwenVL3-30B	27.8	27.6	28.4	27.1	28.5	5.5	22.8	52.4
-Thinking	29.4	**30.6**	31.4	24.1	29.1	5.8	22.8	56.6

Table 5. 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”.

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 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

Model	Indoor Scene Perception				Robot			Grounding		
	Avg.	Static-CC.	Static-IC.	Dynamic	Avg.	Man.	Nav.	Avg.	TG.	TL.
(Sufficient-Coverage) Proprietary										
gpt-4o	29.6	26.3	28.3	**44.6**	**34.8**	**34.0**	35.5	29.9	28.5	**37.0**
o3	38.6	33.9	**43.0**	33.9	33.3	30.9	35.5	34.9	34.9	35.2
o4-mini	34.0	33.3	34.6	33.9	32.8	28.9	**36.5**	29.9	30.2	27.8
gemini-2.5-flash	**39.2**	**40.3**	38.6	38.5	29.9	26.8	32.7	32.5	34.2	24.1
-Thinking	36.5	39.2	35.7	32.3	31.9	26.8	**36.5**	**36.4**	**37.0**	33.3
(Sufficient-Coverage) Open-source										
InternVL2.5-8B	26.6	23.1	27.9	30.8	23.5	30.9	16.8	27.5	26.7	31.5
InternVL3-8B	24.3	21.5	27.9	16.9	29.9	29.9	29.9	24.8	25.3	22.2
InternVideo2.5-8B	27.9	26.9	27.6	32.3	26.5	28.9	24.3	29.0	**29.5**	25.9
QwenVL2.5-7B	27.3	25.3	29.8	23.1	28.9	29.9	28.0	25.4	26.0	22.2
QwenVL2.5-32B	28.7	24.2	**30.5**	33.9	**36.8**	38.1	**35.5**	**29.6**	28.5	**35.2**
QwenVL2.5-72B	**30.6**	**31.7**	27.9	**38.5**	30.4	34.0	27.1	28.4	29.2	24.1
QwenVL3-8B	30.4	30.1	28.7	**38.5**	33.3	**39.2**	28.0	27.5	26.0	**35.2**
QwenVL3-30B	27.1	25.3	26.5	35.4	32.4	33.0	31.8	27.5	26.3	33.3
-Thinking	28.5	25.8	28.7	35.4	31.9	36.1	28.0	27.2	26.3	31.5
(Uniform-50) Proprietary										
claude-haiku-4-5	31.7	34.4	29.4	33.9	28.9	30.9	27.1	29.2	28.1	35.2
gpt-4o	31.6	30.6	31.6	33.9	36.3	**37.1**	35.5	31.6	32.7	25.9
gpt-5	39.0	**39.2**	**39.7**	35.4	33.3	32.0	34.6	34.3	33.5	**38.9**
o3	**39.2**	37.6	**39.7**	**41.5**	**36.8**	32.0	**41.1**	**37.0**	**38.1**	31.5
o4-mini	37.7	38.7	36.8	38.5	32.8	33.0	32.7	33.1	32.4	37.0
gemini-2.5-flash	38.0	37.6	39.3	33.9	31.9	24.7	38.3	31.3	32.0	27.8
-Thinking	36.9	37.1	37.9	32.3	32.4	25.8	38.3	32.8	33.1	31.5
(Uniform-50) Open-source										
InternVL2.5-8B	26.8	22.0	28.7	32.3	32.8	35.0	30.8	28.1	27.1	33.3
InternVL2.5-38B	30.2	26.9	28.3	**47.7**	36.3	36.1	**36.5**	32.2	31.3	**37.0**
InternVL2.5-78B	30.8	28.5	30.9	36.9	29.9	32.0	28.0	30.8	30.2	33.3
InternVL3-8B	26.0	20.4	27.6	35.4	32.8	35.0	30.8	26.6	27.4	22.2
InternVL3-38B	31.0	28.0	33.1	30.8	27.4	27.8	27.1	29.0	29.2	27.8
InternVL3-78B	29.8	24.7	31.6	36.9	**38.7**	43.3	34.6	31.9	32.7	27.8
InternVideo2.5-8B	28.9	25.3	29.4	36.9	24.5	21.6	27.1	31.0	32.0	25.9
LLaVA-Video-7B	30.8	28.5	30.9	36.9	24.0	25.8	22.4	28.1	27.1	33.3
LLaVA-Video-72B	31.2	22.0	**34.6**	43.1	29.4	36.1	23.4	29.0	27.8	35.2
QwenVL2.5-7B	27.0	26.3	27.6	26.1	27.9	28.9	27.1	26.3	25.6	29.6
QwenVL2.5-32B	26.6	22.6	26.8	36.9	34.8	36.1	33.6	28.1	28.5	25.9
QwenVL2.5-72B	29.8	28.5	28.7	38.5	33.8	**45.4**	23.4	**34.0**	**33.8**	35.2
QwenVL3-8B	28.9	22.6	31.6	35.4	30.9	32.0	29.9	30.4	30.2	31.5
QwenVL3-30B	29.4	24.2	31.2	36.9	29.9	33.0	27.1	27.5	26.3	33.3
-Thinking	**31.6**	**30.1**	31.6	35.4	30.9	41.2	21.5	28.7	28.8	27.8

Table 6. 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.

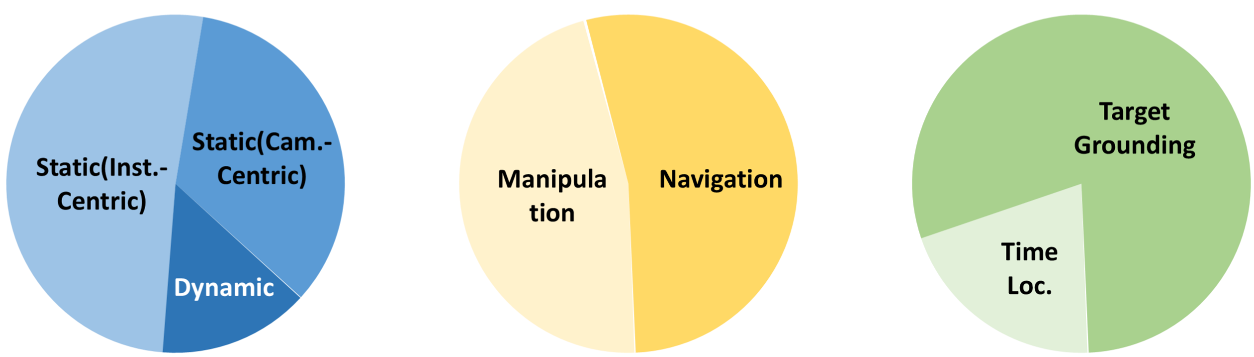

Figure 9. Distribution of subtask proportions across the Indoor Scene Perception Bench, Robot Bench and Grounding Bench.

of Tab.6, indicate that among all models, o3 and gemini-2.5-flash achieve the strongest performance, maintaining the advantage they demonstrated in the main table; notably, o3 excels on instance-centric static scene perception, while gemini-2.5-flash achieves the best performance on camera-centric static scene perception. In addition, InternVL2.5-38B shows 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 from MMSI-Video-Bench. 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.6, InternVL3-78B 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 leads 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, o3 achieves the strongest overall results, demonstrating highly reliable grounding abilities, particularly in identifying target objects within the video. In contrast, gpt-5 excels on temporal localization, showing a clear advantage in accurately pinpointing the relevant time segments.

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.

10.3. Spatially-Grounded Model Evaluation

In recent years, several approaches have emerged that enhance general-purpose models with spatial reasoning capabilities, either by training them on spatial reasoning data (e.g., SpaceQwen [22]) or by introducing architectural modifications to equip models with latent spatial representations (e.g., VLM3R [11], Spatial-MLLM [41]). These methods aim to endow models with spatial intelligence and have reported noticeable improvements on certain spatial reasoning benchmarks. We evaluate these spatially grounded models under our benchmark as well under the Uniform-50 setting. As shown in Tab.7, compared with their respective base models, only SpaceQwen exhibits a slight performance gain, 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 [29, 47]: although such models may perform well on specific spatial reasoning datasets, their capabilities do not generalize effectively to other benchmarks and may even impair previously strong abilities. These findings further highlight the challenge posed by our benchmark and its emphasis on assessing models' spatial reasoning generalization.

11. Preliminary Exploration for Model Improvement

In the main paper, 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.

11.1. 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, enabling it to better perceive global scene geometry. As illustrated in Fig.11, 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.10. Only QwenVL2.5-72B exhibited a slight improvement (2.1%), while the other models showed almost no gains, 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 can handle relatively simple scenes, such as indoor scanning, it often fails in complex scenarios involving multi-room or multi-floor scans or dynamic scenes. In these failure

Model	Modifications	IF. Accuracy	Avg	SC.	MU.	Plan.&Pred.	Cross-Video
(base) LLaVA-Video-7B	-	100%	29.54	30.17	29.55	27.54	30.23
Spatial-MLLM	Architecture	97.6%	23.58	22.41	23.86	25.6	23.84
VLM3R	Architecture	16.1%	4.97	6.47	4.55	4.83	1.74
(base) QwenVL2.5-3B	-	100%	26.92	26.51	27.65	28.99	24.42
SpaceQwen	Fine-tuning	100%	27.46	26.08	31.06	27.05	26.16

Table 7. The table presents the performance of various spatially grounded models and their corresponding base models across the major task categories of MMSI-Video-Bench. "IF" denotes Instruction Following, while "SC," "MU," "Plan," and "Pred" refer to Spatial Construction, Motion Understanding, Planning, and Prediction, respectively.

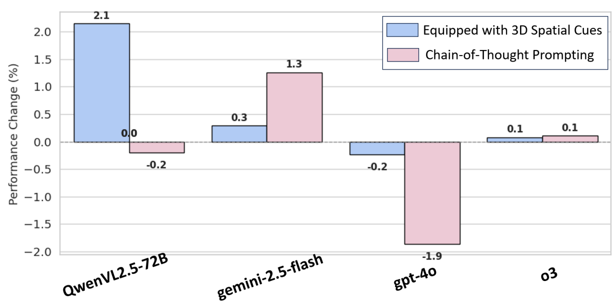

Figure 10. Effect of different methods on model performance.

cases, the rendered images provide little to no useful information for the models and may even introduce noise. This reflects an inherent limitation of VGGT; to consistently provide accurate 3D spatial cues, more robust and generalizable tools are needed.
- Issues in utilizing 3D spatial cues. Examination of the models' reasoning processes revealed that the models fail to effectively leverage the 3D spatial cues. In many cases, the cues are either ignored or not correctly associated with the video content and the question, even though our prompts explicitly instructed the model to use them. This indicates that designing spatial cues that are easily interpretable by the models remains an open challenge.

11.2. 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.10, simply encouraging the model to "think step by step" does not consistently improve performance. This aligns with previous findings [46, 47]. 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.

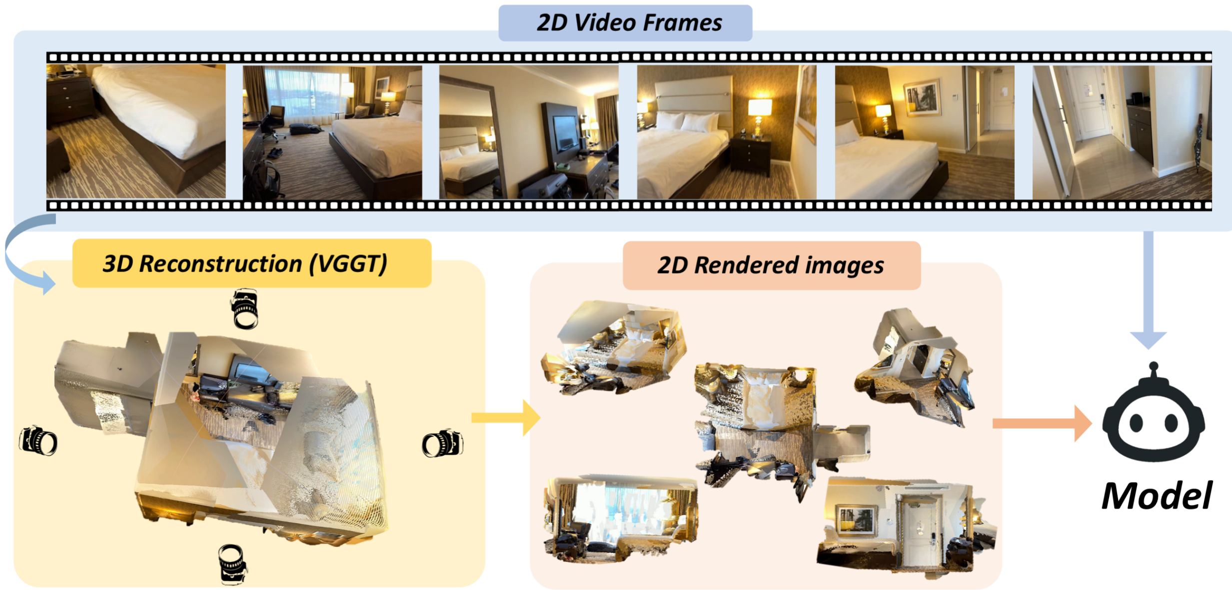

Figure 11. Pipeline of equipping the model with 3D spatial cues.

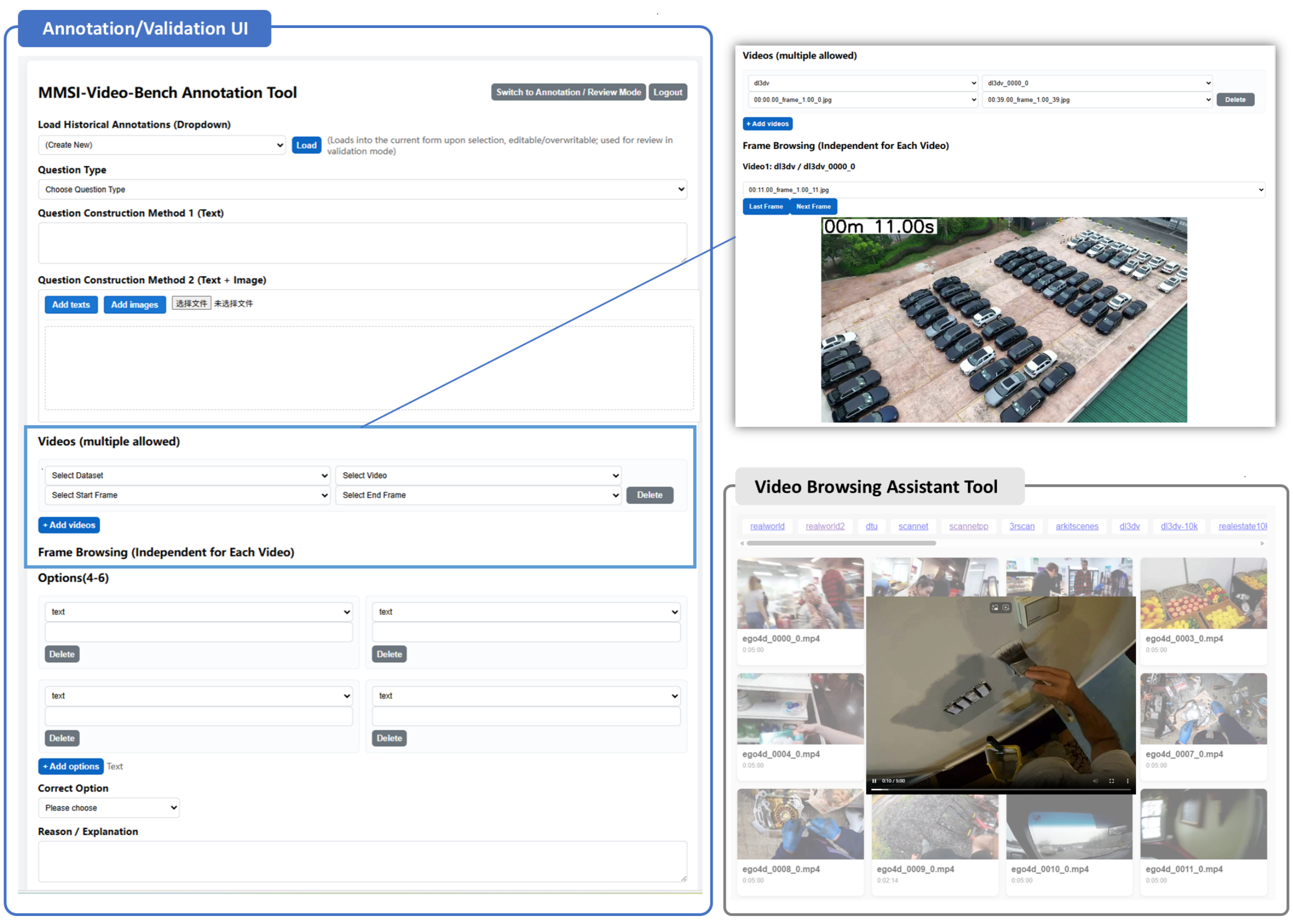

Figure 12. User interface for annotation and quality validation in MMSI-Video-Bench.

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 13. Task Categories and Subtypes in MMSI-Video-Bench. ”Inst.” denotes ”instance”, ”Cam.” denotes ”camera” and ”Rel.” denotes ”relationship”.

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 14. Structure of system and user prompts used in the experiments.