Seeing Together: Multi-Robot Cooperative Egocentric Spatial Reasoning with Multimodal Large Language Models

Kunyu Peng[1,2,*] **Zhikun Zhou**[3] **Kailun Yang**[3,*] **Di Wen**[1] **Ruiping Liu**[1]
Yufan Chen[1,4] **Junwei Zheng**[1,6] **Hao Shi**[5,7] **Yi Zhou**[3]
M. Saquib Sarfraz[8] **Danda Pani Paudel**[2] **Luc Van Gool**[2]
[1]Karlsruhe Institute of Technology [2]INSAIT, Sofia University "St. Kliment Ohridski"
[3]Hunan University [4]University of Oxford [5]Zhejiang University
[6]ETH Zurich [7]Ant Group [8]Mercedes-Benz Tech Innovation

Abstract

Multimodal Large Language Models (MLLMs) have made substantial progress in egocentric video understanding, but their ability to reason cooperatively from multiple embodied viewpoints remains largely unexplored. We study this problem through *multi-robot cooperative dynamic spatial reasoning*, where a model must answer spatial, temporal, visibility, and coordination questions by integrating synchronized egocentric videos from a team of moving robots. To support this setting, we introduce **CoopSR**, the first benchmark for this task, together with **EgoTeam**, a multi-robot egocentric QA dataset. EgoTeam contains $114,227$ QA pairs spanning 19 question types, four difficulty tiers, and three team sizes in Habitat and iGibson, along with a real-world test set of around $2,326$ QAs collected using two quadruped robots. We further propose **SG-CoR** (*Spectral and Geometry-Aware Cooperative Reasoner*), an MLLM framework for fine-grained cooperative spatial reasoning. SG-CoR combines dynamics-aware multi-robot frame sampling, spectral- and geometric-guided view fusion, and geometry- and motion-informed prompt distillation, enabling the model to benefit from privileged robot-pose supervision during training while requiring only egocentric videos at test time. Across 22 MLLM baselines, SG-CoR consistently improves cooperative reasoning, outperforming the strongest fine-tuned baseline by **+3.87%** on Habitat and **+7.12%** on iGibson. It also shows stronger generalization to unseen team sizes and real-world robot tests. Code can be found at `https://github.com/KPeng9510/seeing-together.git`.

1 Introduction

Embodied AI is rapidly shifting from isolated agents to teams of robots operating in shared physical spaces, including warehouse fleets, search-and-rescue squads, inspection drones, and household assistants [63, 18, 10, 29]. In these settings, no single robot has a complete view of the world. Each agent observes only a partial, egocentric, and often noisy slice of the scene. Answering even a seemingly simple operator query, such as "Which robot is pushing the object?" or "Can any robot see the blocked doorway?", may require integrating observations across robots, grounding objects across viewpoints, reasoning about inter-robot spatial relations, and building a team-level understanding of the scene. Beyond perception alone, real-world deployment also requires robots to coordinate tightly during physically coupled tasks, such as two robots jointly carrying a long object through cluttered spaces, where failures in shared spatial understanding can lead to collisions, blockage, or

*Correspondence: kunyu.peng@kit.edu and kailun.yang@hnu.edu.cn

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

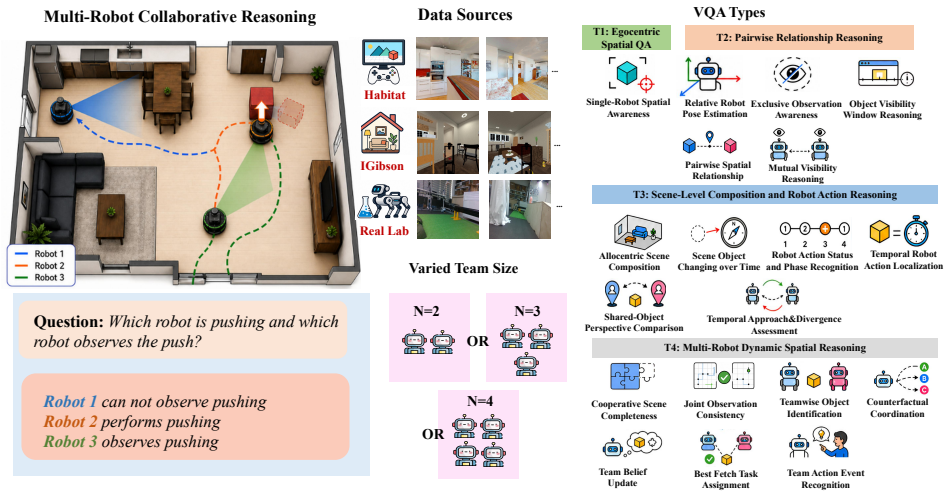


Figure 1: Overview of the **CoopSR** benchmark. It evaluates spatial, temporal, visibility, and coordination reasoning using synchronized egocentric views from variable-size robot teams in simulated and real-world environments.

unsafe interactions. We call this capability *cooperative spatial reasoning*: reasoning over multi-robot egocentric video streams to build a shared understanding of a dynamic environment.

Despite its importance, this setting remains largely unexplored in current Multimodal Large Language Models (MLLMs). Existing MLLMs [68, 63, 72, 24] are primarily trained and evaluated on single-view images or videos, often from a single human or robot perspective. They typically treat visual tokens as a flat sequence and lack mechanisms for explicitly aligning, filtering, and fusing observations from multiple embodied agents. A naive extension—simply concatenating N robot video streams—is not scalable: the number of visual tokens grows rapidly with team size, while adjacent frames and overlapping viewpoints introduce substantial redundancy. More importantly, concatenation alone does not teach the model *how* to reason collaboratively across agents, such as determining which robot has the best viewpoint, resolving contradictory observations, or inferring scene-level structure from distributed egocentric evidence.

Existing benchmarks [16, 27, 50] do not adequately address this problem. Egocentric VQA benchmarks [50, 12] mainly consider a single observer, while EgoExo QA benchmarks [28, 21] focus on reasoning between a fixed exocentric view and one moving egocentric view. Recent VQA benchmarks for moving egocentric agents, *e.g.*, MA-EgoQA [29], assume that the agent already possesses sufficient spatial reasoning ability in dynamic scenes and primarily emphasize lifelong human behavior understanding, rather than fine-grained joint spatial reasoning, robot action understanding, and environment awareness. In contrast, our benchmark is uniquely designed to evaluate whether MLLMs can understand a shared dynamic environment through the eyes of a robot team. It concentrates on cross-view grounding, inter-robot spatial relations, team-level scene understanding, and fills a critical gap by directly assessing team-level egocentric perception and reasoning for multi-robot embodied intelligence.

To fill this gap, we introduce **CoopSR**, a benchmark for *cooperative multi-robot visual question answering*, and **EgoTeam** dataset collected from both simulators and real-world labs. **EgoTeam** contains $114K$ QA pairs covering 19 QA types along five axes: spatial, exploration, relational, robot action, and visibility reasoning. The questions span four levels, from egocentric scene understanding to cooperative scene perception, and three team sizes in habitat, iGibson, and real-world environments. **CoopSR** evaluates 22 MLLMs on synchronized egocentric observations from variable-size robot teams, targeting key skills such as spatial awareness, robot-object relations, scene composition, temporal robot action reasoning, mutual visibility, shared-object perception, team belief updates, and counterfactual coordination. Unlike prior VQA benchmarks focused on single-agent perception, **CoopSR** evaluates whether models can integrate partial multi-robot viewpoints, reason about complementary observations and inter-robot relations, and infer coherent scene-level answers from distributed embodied evidence.

However, existing models [26, 68, 72, 24, 1, 59] perform reasonably well on single-view reasoning [12, 50] or EgoExo setting [23] but degrade sharply on questions requiring inter-robot integration,

cross-team-size generalization, or conflicting viewpoints. This failure pattern suggests a structural bottleneck: current MLLMs lack inductive biases for agent identity, geometric co-visibility, and multi-robot egocentric view fusion. Motivated by these findings, we propose **SG-CoR** (*Spectral and Geometry-Informed Cooperative Reasoner*), an MLLM architecture designed for cooperative spatial reasoning. **SG-CoR** has three key components. First, a training-free two-stage sampler compresses each robot stream by jointly selecting the most query-relevant frames and refining them using temporal frequency energy ranking over visual features. Second, a spectral and geometry-informed attentional block fuses robot observations while injecting a geometry-prior from the data-rich simulator and the spectral cues into attentional fusion. Ground-truth poses are used only during training, and deployment requires no external localization oracle. Third, a pose-aware teacher distills physical priors into the student's prefix tokens, aligning the language prompt with the pose prior provided by the simulator. The student prompt is utilized as geometric prior during deployment, as the real world cannot be guaranteed to provide precise pose cues.

On the **CoopSR** benchmark, **SG-CoR** consistently outperforms zero-shot, SFT, RAG [29, 26], and key frame selection [56, 57] baselines across different evaluation settings. The best **SG-CoR** variant, using a Qwen2.5-VL [2] backbone, achieves an average accuracy of 70.55% on Habitat [48] and 70.82% on iGibson [32], improving over the strongest baseline by $+3.87\%$ and $+7.12\%$, respectively. **SG-CoR** also shows superior cross-MLLM-backbone generalizability. These results indicate that **SG-CoR** benefits from an explicit cooperative reasoning design. Further evaluation on cross-team-size and sim-to-real generalizability further demonstrates the superiority of our proposed method. In summary, our contributions are as follows:

- We introduce **cooperative spatial reasoning** as a new embodied AI problem, where an MLLM must answer language queries by integrating synchronized egocentric observations from multiple moving robots.
- We present **EgoTeam**, a dataset for cooperative multi-robot visual question answering, containing $> 114K$ question-answer pairs across 19 QA types in simulated and real-world environments. We conduct a comprehensive evaluation of 22 MLLM baselines on this dataset and build this first **CoopSR** benchmark.
- We propose **SG-CoR**, a cooperative MLLM framework that introduces a spectral energy-aware multi-robot relevant frame sampler, spectral and geometry- and motion-informed multi-robot fusion, and geometry- and motion-informed prompt-space distillation for scalable reasoning over multi-robot egocentric videos without requiring pose data during test time. **SG-CoR** delivers state-of-the-art performance across Habitat, iGibson, cross-team-size, and sim-to-real settings.

2 Related Work

Multi-agent perception and cooperation. Existing works on multiple embodied agents study cooperation across diverse environments [39, 62, 73, 6, 49, 71, 53], with systems such as CoELA [70], Co-NavGPT [65], OmniVLN [38], DeCoNav [74], AirCopBench [67], TZPP [61], COHERENT [36], and CoMaTrack [37] using LLMs/VLMs for multi-agent planning, coordination, and exploration. Other studies address more realistic constraints such as limited communication and partial observability [66, 25], while structured or leadership-based prompting has been shown to improve teamwork efficiency [20]. PARTNR [8] further provides a large-scale human-robot collaboration benchmark.

Multimodal large language models. The past two years have seen rapid progress in MLLMs that combine vision encoders with LLM backbones: Video-LLaVA [33], InternVL [11], Qwen2.5-VL [2], Qwen3-VL [1], LLaVA-Next-Video [72], and VideoLLaMA3 [68]. These models are trained on single-image or short-video data and treat all visual tokens uniformly, with no mechanism for agent-level identity or inter-agent geometric constraints. We systematically evaluated 22 MLLMs, including zero-shot MLLMs, supervised fine-tuned MLLMs, Retrieval-Augmented-Generation methods [29, 26], and key frame selection methods [57, 56]. Existing MLLMs lack cross-robot alignment, multi-view fusion, and team-level belief modeling, therefore, we propose **SG-CoR**, which integrates dynamics-aware sampling, geometry-guided fusion, and geometry- and motion-informed prompt distillation for cooperative spatial reasoning.

Benchmarks of egocentric video question answering. Early egocentric vision research [15, 17, 30, 52, 3, 43, 45, 35] was driven by foundational but relatively small datasets such as ADL [46]. The field later expanded through large-scale benchmarks including EPIC-KITCHENS [13], Ego4D [18], and

Table 1: Comparison of our **EgoTeam** with other egocentric QA datasets.

Dataset	Source	#QAs	Size (hrs)	#Clips	Dur./Clip	MultiView	Inter Robot Dynamic	Robot Pose
EgoSchema [42]	Ego4D [18]	5K	250	5,063	3 min	✗	✗	✗
EgoPlan-Bench [10]	Ego4D [18] & EPIC-KITCHENS [13]	4.9K	-	4,939	-	✗	✗	✗
EgoThink [12]	Ego4D [18]	0.7K	-	595	-	✗	✗	✗
EgoMemoria [64]	Ego4D [18]	7K	-	629	30 s to 1 h	✗	✗	✗
HourVideo [7]	Ego4D [18]	12.9K	381	500	20 min to 2 h	✗	✗	✗
EgoLifeQA [63]	EgoLife [63]	3K	266	6	44.3 h	✓	✗	✗
EgoTeam (ours)	-	114K	385.7	21,124	1 min to 6 min	✓	✓	✓

EPFL-Smart-Kitchen [5], which broadened egocentric video understanding to diverse daily activities. Subsequent datasets targeted more specialized settings, including procedure learning [4, 54, 75], collaborative assistance [60], and joint egocentric-exocentric understanding [19, 23]. Building on these resources, recent benchmarks have advanced temporal reasoning, forecasting, and planning in first-person video [34, 14, 47, 58, 42, 31]. Zhang *et al.* [69] propose pairwise egocentric video question answering in day and night time. More recent efforts further explore long-term video understanding [64, 7]. Jung *et al.* [28] and He *et al.* [21] focus on egocentric and exocentric videos for cross-view video understanding. MA-EgoQA [29] focuses on long-term human behavior reasoning using multiple agents, which focuses on coarse-level QAs. Unlike these benchmarks, our **CoopSR** directly targets fine-grained cooperative spatial reasoning from synchronized multi-robot egocentric videos. It evaluates whether MLLMs can integrate partial and dynamic robot viewpoints to reason about inter-robot relations, shared scene structure, visibility, and team-level coordination.

3 CoopSR-Bench: Benchmark and Datasets

Data collection and annotation. Metadata collection. We first collect multi-robot egocentric data in Habitat [48] and iGibson [32]. For each scene, we deploy teams of 2, 3, and 4 robots and generate 15 exploration episodes with diverse trajectories and robot-object interaction events, such as object pushing, yielding diverse layouts, viewpoints, and temporal dynamics. For each team size $N \in \{2, 3, 4\}$, episodes are collected in separate simulator runs under a common collection protocol and comparable scene distributions. Robot trajectories are generated jointly for each team and are not paired across team sizes. For each episode, we record robot poses, pairwise relative poses, egocentric RGB-D videos, semantic segmentation videos, and object to object/robot spatial relationships per timestep. These annotations support QA construction over egocentric perception, cross-view association, temporal reasoning, and multi-robot coordination. We also collect a real-world test set by exploring a room with two quadruped robots and using motion capture systems to provide pose metadata for QA creation; details are provided in the appendix.

QA annotations. For iGibson [32] and Habitat [48], we use simulator-recorded scene attributes as metadata for QA generation and then rephrase the QAs by GPT-4o [44]. QA pairs are derived from the aforementioned extracted metadata. This enables scalable QA generation while ensuring that each question is grounded in the scene state and has an unambiguous metadata-derived answer. For the real-world quadruped robot dataset, we manually annotate scene captions as metadata at adaptive intervals of fewer than 30 frames to avoid obvious scene changes within each interval, and then create QAs based on metadata. In total, we collect $1.75h$ of real-world data across 13 scene settings. Additional statistics and comparisons with existing egocentric datasets are shown in Table 1. To improve evaluation reliability, four annotators manually review QA pairs in the validation and test splits to ensure correctness with a majority agreement ratio of 0.9.

QA taxonomy. Our QA taxonomy evaluates multi-robot egocentric reasoning across four hierarchical reasoning tiers of increasing cooperative scope, ranging from local spatial perception to team-level coordination, as illustrated in Figure 1.

T1: Egocentric spatial QA. T1 level evaluates spatial understanding from a single robot's first-person observation. It includes single-robot spatial awareness, such as recognizing object locations, directions, distances, and layouts.

T2: Pairwise relationship Reasoning. T2 level focuses on relationships between two entities, including robots, objects, and viewpoints. It evaluates whether a model can identify information visible to only one robot, reason about object visibility and occlusion over time, determine pairwise spatial relations such as left/right, front/behind, near/far, and facing direction, and assess whether two robots can see each other or share overlapping fields of view.

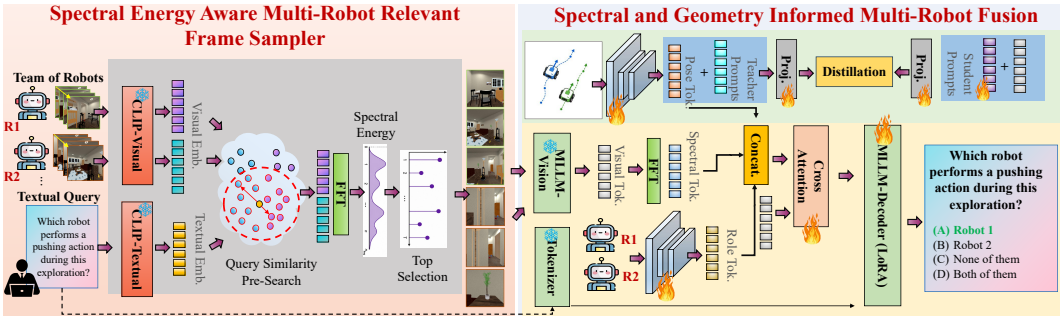


Figure 2: Overview of **SG-CoR**, which combines query-guided spectral frame sampling, spectral- and geometry-informed multi-robot fusion, and prompt-space distillation for cooperative spatial reasoning on **CoopSR**.

T3: Scene-level composition and robot action reasoning. T3 level requires integrating multiple egocentric observations into a coherent scene-level representation. It covers allocentric scene reconstruction, temporal tracking of object changes, recognition of robot action phases, estimation of action duration, comparison of shared objects across viewpoints, and reasoning about whether robots approach or move away from one another or from target objects over time.

T4: Multi-robot dynamic spatial reasoning. T4 level evaluates high-level collaborative reasoning over the full robot team. It tests whether the model can assess collective scene coverage, detect consistency or conflict among robot observations, identify the same object across views, reason counterfactually about alternative coordination strategies, update team-level beliefs from new observations, assign fetch tasks based on visibility, distance, and action context, and infer the overall team event, including acting and observing robots. Overall, the taxonomy provides a structured evaluation of multi-robot egocentric reasoning, progressing from single-view perception to pairwise relations, global scene understanding, and dynamic team coordination.

Baselines. Zero-shot MLLMs: We evaluate representative image/video MLLMs, including Qwen2.5-VL-7B/72B [2], Qwen3-VL-8B [1], InternVL3-8B [59], LLaVA-NeXT-Video-7B [72], and VideoLLaMA3-7B/2B [68], to assess their zero-shot performance on **CoopSR**. Supervised fine-tuning baselines: We fine-tune the same major backbones with Low Rank Adaptation (LoRA) [22] on the MLLM decoder on **CoopSR** to test whether standard task adaptation is sufficient and to isolate the architectural gains of **SG-CoR**. Retrieval-augmented baselines: We compare with VideoRAG [41] and EgoRAG [63] to evaluate whether retrieved video or egocentric evidence improves reasoning over long, redundant multi-robot streams. Keyframe-selection baselines: We include TSPO [57] and KTV [56] under the best model from Supervised Fine-Tuning (SFT) to examine the performance of keyframe selection approaches. Together, these baselines control for *model capability*, *task adaptation*, and *evidence selection*, allowing us to verify that **SG-CoR** improves **CoopSR** through cooperative inductive biases for agent identity, co-visibility, temporal dynamics, and multi-robot fusion.

Dataset splits and evaluation metric. Data splits. For the dataset with $114,227$ QAs collected from Habitat [48] and iGibson [32], we use $119/12/23$ unique scenes for train/val/test for Habitat, while the iGibson counterpart uses $8/2/4$ scenes for train/val/test. All 13 scenes of the real-world set are used as the test set, which contributes to $2,326$ QAs in total. Multiple-choice QA (MC4). For all 19 QA types, models produce a single letter (A/B/C/D) and are scored Primary metrics are MC4 Accuracy (overall and per task-difficulty/team-size level), including per-difficulty-tier accuracy, per-source accuracy (non-iGibson *vs.* iGibson), and N-robot generalization (train on $N = [2, 3]$, evaluate on $N = 4$). AVG denotes the averaged QA performance on all samples without grouping.

4 Methodology

4.1 Problem Formulation

We study *multi-robot egocentric cooperative spatial reasoning*, where N robots jointly answer a query from synchronized egocentric video streams. Robot r observes $\mathcal{V}_r = \{\mathbf{x}_{r,t}\}_{t=1}^{T}$ and receives controls $\mathbf{u}_{r,t}$. During training, we additionally use pose estimates $\hat{\mathbf{u}}_{r,t} = (\hat{\mathbf{p}}_{r,t}, \hat{\mathbf{e}}_{r,t}) \in \mathrm{SE}(3)$ from

simulators or motion capture; these poses are unavailable at test time. Given $\mathcal{V} = \{\mathcal{V}_r\}_{r=1}^{N}$ and query $\mathbf{q}$, the model predicts $\mathbf{y} = f_\theta(\mathcal{V}, \{\hat{\mathbf{u}}_{r,t}\}, \mathbf{q})$ during training, $\mathbf{y} = f_\theta(\mathcal{V}, \mathbf{q})$ during testing. T and N denote number of frames and robots. The active robot count N and episode-local robot indices are provided as input metadata. The robot index is passed to the model associated with its video stream. This setting requires selecting sparse evidence across robots and time, resolving redundant or conflicting viewpoints, and reasoning over shared physical constraints.

4.2 Spectral and Geometry-Aware Cooperative Reasoner (SG-CoR)

Current MLLMs struggle with multi-robot cooperative spatial reasoning due to limited cross-robot dynamics modeling and weak use of physical priors. We propose **SG-CoR** (*Spectral and Geometry-Aware Cooperative Reasoner*), which combines spectral energy-aware frame sampling with geometry-informed view fusion and prompt distillation. It selects informative multi-robot frames, injects motion and spatial priors during training, and enables pose-free cooperative reasoning at inference. The whole pipeline of our approach is shown in Figure 2.

Spectral Energy-aware Multi-Robot Relevant Frame Sampler (SE-MR^{2}FS). Feeding all frames from all robots yields $\mathcal{O}(NT)$ visual tokens for N robots and T frames, which is costly and redundant due to adjacent near-duplicate frames and overlapping robot views. We independently compress each active robot stream to $K \ll T$ frames using a training-free two-stage sampler: a frozen vision–language model first retrieves query-relevant candidates, which are then refined by temporal frequency energy over visual features. Specifically, each frame is embedded with a frozen CLIP [51] image encoder ϕ_v as $\mathbf{z}_{r,t} = \phi_v(\mathbf{x}_{r,t}) \in \mathbb{R}^{d_c}$, while the query is encoded by the corresponding CLIP text encoder as $\mathbf{q} \in \mathbb{R}^{d_c}$. Both embeddings are ℓ_2-normalized.

CLIP-based semantic candidate selection. For each robot r, we score every frame by its query similarity according to Eq. 1.

$$\mathbf{s}_{r,t} \;=\; \langle \mathbf{z}_{r,t},\, \mathbf{q} \rangle \;\in\; [-1, 1]\,, \tag{1}$$

and retain the top-M frames as a candidate set as shown in Eq. 2.

$$\mathcal{C}_r \;=\; \mathrm{TopM}\big(\{\mathbf{s}_{r,t}\}_{t=1}^{T};\; M\big). \tag{2}$$

This step removes frames that are obviously irrelevant to the question before any further step.

Temporal FFT energy on visual features. Many questions depend on understanding temporal dynamics, such as which robot moved first, when an object was placed, or which robot interacted with another object or agent. These changes cannot be captured reliably by static CLIP [51] similarity alone. We therefore compute a per-frame motion descriptor by applying a discrete Fourier transform along the temporal axis of CLIP visual features. For a sliding window $\mathcal{W}(t)$ of length w centered at t, let $\mathbf{Z}_{r,\mathcal{W}(t)} \in \mathbb{R}^{w \times d_c}$ denote the stacked CLIP features after subtraction of the local mean. Fourier transformation is conducted according to Eq. 3.

$$\Phi_{r,t}[k] \;=\; \log\!\Big(1 + \tfrac{1}{d_c}\big\|\mathcal{F}_w\big[\mathbf{Z}_{r,\mathcal{W}(t)}\big]_k\big\|_2^2\Big), \qquad k = [1, \ldots, B]\,, \tag{3}$$

where $\mathcal{F}_w$ is the length-w rFFT, the zero-frequency component is discarded so that the descriptor responds only to change, and the remaining bins are pooled into B bands. The scalar frequency energy of frame (r, t) is the sum of retained bands according to Eq. 4.

$$\mathbf{e}_{r,t} \;=\; \sum_{k=1}^{B} \Phi_{r,t}[k]. \tag{4}$$

Because the FFT operates on CLIP embeddings rather than pixels, $\mathbf{e}_{r,t}$ captures semantically meaningful motion while remaining insensitive to nuisance pixel-level perturbation. We restrict the FFT energy to the candidate set $\mathcal{C}_r$ and standardize it within that set to get $\tilde{\mathbf{e}}_{r,t}$. The final per-frame score combines frequency energy with an optional semantic-refinement term according to Eq. 5,

$$\rho_{r,t} \;=\; \tilde{\mathbf{e}}_{r,t} \;+\; \tilde{\mathbf{s}}_{r,t}, \qquad t \in \mathcal{C}_r, \tag{5}$$

and the K frames with highest $\rho_{r,t}$ within $\mathcal{C}_r$ are retained.

Spectral token extraction. The same FFT processing yields a compact robot-level descriptor that is later injected into the fusion stage. After frame sampling, each robot supplies K tunable visual

tokens $\{\mathbf{h}_{r,k}\}_{k=1}^{K}$ produced by MLLM visual encoder. We compute the magnitude spectrum of this short sequence, retain its low-frequency coefficients, and project them through a small MLP with LayerNorm to obtain $\mathbf{F}_r \in \mathbb{R}^{d_F}$. $\mathbf{F}_r$ encodes how each robot's evidence evolves over the sampled subset, and is consumed by the fusion module as a per-robot temporal summary.

Spectral and Geometry-Informed Multi-Robot Fusion (SGI-MRF). Single-robot video understanding mainly relies on temporal context within one coherent viewpoint. In contrast, multi-robot reasoning requires integrating geometrically constrained and mutually dependent observations. Robots may view the same object from different angles, suffer different occlusions, or share overlapping fields of view, so the model must reason about co-visibility, relative pose, and observation reliability.

Naive fusion methods such as concatenation, addition, or self-attention over pooled tokens treat robot observations as independent and interchangeable. They therefore cannot reliably resolve viewpoint conflicts, infer shared scene structure, or update team-level beliefs when observations disagree. SGI-MRF is designed to explicitly model these cooperative dependencies.

Given the sampled visual tokens $\{\mathbf{h}_{r,k}\}_{k=1}^{K}$, we construct a robot-level visual state μ_r and a learned uncertainty vector σ_r to support reliability-aware fusion. To capture temporal dynamics, we compute a Fourier embedding $\mathbf{F}_r$ from the sampled tokens, and encode robot motion commands into a pose embedding $\mathbf{P}_r$. We then form a geometric- and motion-aware robot token by combining pose, spectral, and robot-role embeddings as Eq. 6, using linear projection layers $\mathbf{W}_p$ and $\mathbf{W}_f$. Here, $\mathbf{E}_r$ is a learned robot-role embedding indexed by the episode-local identity of Robot r. It provides an explicit robot-identity cue and is added to the projected motion and spectral embeddings to form the auxiliary token $\mathbf{A}_r$. Robot indices are local to each episode and do not identify a fixed physical robot across episodes.

$$\mathbf{A}_r = \mathbf{W}_p\mathbf{P}_r + \mathbf{W}_f\mathbf{F}_r + \mathbf{E}_r. \tag{6}$$

The visual states attend to these auxiliary tokens through cross-attention [9] according to Eq. 7.

$$\hat{\mu}_{1:N} = \mathrm{CrossAttention}(\mu_{1:N}, \mathbf{A}_{1:N}, \mathbf{A}_{1:N}), \tag{7}$$

allowing each robot's view to be updated with motion, identity, and temporal cues.

The uncertainty used for pooling is estimated from sampled visual features, rather than simulator metadata or robot poses. Let $\mathbf{h}_{r,k}$ denote the projected feature of the k-th sampled frame from robot r. A learned temporal query produces nonnegative attention weights $\mathbf{a}_{r,k}$ satisfying $\sum_{k=1}^{K}\mathbf{a}_{r,k} = 1$. We compute the attention-weighted visual mean and temporal variance as

$$\mu_r^{\mathrm{vis}} = \sum_{k=1}^{K}\mathbf{a}_{r,k}\mathbf{h}_{r,k}, \mathbf{v}_r = \sum_{k=1}^{K}\mathbf{a}_{r,k}\left(\mathbf{h}_{r,k} - \mu_r^{\mathrm{vis}}\right)^2. \tag{8}$$

The initial visual uncertainty is

$$\sigma_r^{\mathrm{vis}} = \mathrm{Softplus}\left(\mathbf{g}_\sigma\left(\mu_r^{\mathrm{vis}}\right)\right) + \mathrm{sg}\left(\sqrt{\mathbf{v}_r}\right), \tag{9}$$

where $\mathbf{g}_\sigma$ is a learned uncertainty head, sg denotes stop-gradient, and the square, square root, and Softplus operations are elementwise. The learned component models sample-dependent uncertainty, while the empirical component measures temporal disagreement among the selected frames.

SGI-MRF further projects this uncertainty and incorporates a spectral contribution:

$$\bar{\sigma}_r = \mathrm{Softplus}\left(\mathbf{W}_\sigma\sigma_r^{\mathrm{vis}}\right), \sigma_r = \bar{\sigma}_r + 0.1\,\tilde{\mathbf{e}}_r\mathbf{1}, \tag{10}$$

where $\mathbf{W}_\sigma$ is a learned projection, $\tilde{\mathbf{e}}_r$ denotes the normalized norm of the robot-level Fourier embedding F_r, and $\mathbf{1}$ has the same dimension as $\bar{\sigma}_r$.

The reliability weights underlying are

$$\mathbf{w}_r = \frac{\exp\left(-\|\sigma_r\|_2\right)}{\sum_{j=1}^{N}\exp\left(-\|\sigma_j\|_2\right)}. \tag{11}$$

The implementation additionally applies norm clipping for numerical stability. These weights constitute a learned reliability-aware pooling mechanism, rather than a closed-form Bayesian posterior.

The fused visual states are aggregated into the global team belief using these weights and the projection $\mathbf{W}_b$, as shown in Eq. 12.

$$\mathbf{z} = \mathbf{W}_b \left(\sum_{r=1}^{N} \mathbf{w}_r \hat{\mu}_r \right). \tag{12}$$

Geometry- and Motion-Informed Prompt-Space Distillation (GMIPSD). We propose GMIPSD, a privileged-information learning framework that uses simulator-provided ground-truth robot poses during training but requires no pose input at inference. We construct geometry- and motion-informed teacher prompts from robot poses to encode each robot's state and inter-robot spatial configuration. Learnable student prompts are then trained to align with these teacher prompts, distilling pose-induced physical awareness into task-adaptive latent representations. In inference, only the student prompts are used, enabling geometry- and motion-informed multi-robot fusion from egocentric videos alone. The fused vector $\mathbf{z}$ conditions the language backbone LM_ϕ via projected prefix tokens, which are physically aligned through distillation from a trajectory-aware teacher used only at training time. Both teacher and student output L prompt vectors of dimension d, add learned positional codes, and apply LayerNorm. The teacher uses privileged inputs, *i.e.*, ground-truth pose embeddings $\bar{\mathbf{p}}^{\mathrm{emb}}$ and learnable prefix tokens $\bar{\mathbf{C}}$ of the teacher, to emit a physically informed prompt as Eq. 13.

$$\mathbf{P}_T \;=\; \mathbf{g}_T\big(\bar{\mathbf{p}}^{\mathrm{emb}}, \bar{\mathbf{C}}\big) \;\in\; \mathbb{R}^{L\times d}. \tag{13}$$

The student counterpart has access only to test-time quantities, *i.e.*, the global belief $\mathbf{z}$ and emits $\mathbf{P}_S = \mathbf{g}_S(\mathbf{z}, \mathbf{C})$, where $\mathbf{C}$ are learnable student prompts. Here, $\mathbf{C}$ denotes learnable student prompt tokens, whereas $\mathbf{C}_S \in \mathbb{R}^{L\times d}$ denotes the learned student positional embeddings shown as grey blocks in Figure 2. Writing the prompt generator before positional addition and normalization as g_S^{raw}, the student output is

$$\mathbf{P}_S = \mathrm{LayerNorm}\left(g_S^{\mathrm{raw}}(\mathbf{z}, \mathbf{C}) + \mathbf{C}_S\right). \tag{14}$$

Thus, $\mathbf{g}_S$ includes both positional addition and LayerNorm. The projected student prompts $\mathbf{W}_P\mathbf{P}_S$ are concatenated with the projected global team belief $\mathbf{W}_z\mathbf{z}$ to form the decoder prefix. Teacher prompts $\mathbf{P}_T$ provide the distillation target and are not included in the decoder prefix. The $\mathbf{P}_S$ and $\mathbf{P}_T$ are aligned with a distillation loss $L_{distill}$ formulated by cosine similarity distance. In inference, the teacher is discarded; the student alone produces the prompt that conditions the language model, as shown in Eq. 15.

$$\tilde{\mathbf{H}} \;=\; \big[\,\mathbf{W}_z\mathbf{z};\; \mathbf{W}_P\mathbf{P}_S;\; \mathrm{Embed}(\hat{\mathbf{q}})\,\big], \qquad \mathbf{y} \;=\; \mathrm{LM}_\phi(\tilde{\mathbf{H}}), \tag{15}$$

where $\mathbf{W}_z$ and $\mathbf{W}_P$ are linear projection layers. Embed indicates the tokenizer and $\hat{\mathbf{q}}$ denotes original query. Because distillation happens in prompt space rather than output space, the student is supervised to recover the physical latent structure that would have best conditioned the LM, not merely its final answer – this proves markedly more sample-efficient than standard response distillation.

Training objective. We train end-to-end with a composite objective:

$$\mathcal{L} \;=\; \lambda_{\mathrm{LM}}\mathcal{L}_{\mathrm{LM}} \;+\; \lambda_d\,\mathcal{L}_{\mathrm{distill}}, \tag{16}$$

where $\mathcal{L}_{\mathrm{LM}}$ is the next-token cross-entropy on $\mathbf{y}$ and $\mathcal{L}_{\mathrm{distill}}$ is the geometry- and motion-informed prompt distillation loss. The loss weights λ_{LM} and λ_d are selected using grid search.

5 Experimental Results

Implementation details. All experiments are conducted on 4 NVIDIA A100 GPUs. We precompute CLIP and Qwen visual features before training. The model is trained for 10 epochs using AdamW [40] with learning rate 1×10^{-4}, per-GPU batch size 16, gradient accumulation 4, and gradient clipping 1.0. Learnable prompts number is 8, M, w and K are selected as 32, 4 and 8, LoRA fine-tuning on MLLM decoder is applied with rank 8, $\alpha = 16$, and dropout 0.05. λ_{LM} and λ_d are selected as 1.0 and 0.3. Hyperparameters are selected through grid searching and validation set performance. The number of trainable parameters of **SG-CoR** with Qwen2.5-VL-7B [24] is 23.8M.

Analysis of the benchmark. Table 2 compares different model groups on **CoopSR** across Habitat [48] and iGibson [32], covering four reasoning tiers (T1–T4) and team sizes $N \in \{2, 3, 4\}$. Zero-shot MLLMs obtain low average accuracies, mostly between 31% and 42%, showing that

Table 2: CoopSR benchmark results. We report accuracy for four reasoning tiers (T1–T4), different team sizes ($N = 2, 3, 4$), and the overall average. Note: AVG is computed over all QA samples, rather than by averaging across task types or robot-team sizes. Team-size results compare independently generated episode sets with matched collection protocols; they do not constitute a trajectory-controlled comparison of adding robots.

Datasets / Models	Habitat [48] T1	T2	T3	T4	N=2	N=3	N=4	AVG	iGibson [32] T1	T2	T3	T4	N=2	N=3	N=4	AVG
Zero-shot evaluations																
Qwen2.5-VL-7B-Instruct [2]	23.57	39.61	28.40	34.35	34.66	34.99	31.03	33.51	39.44	50.89	28.12	29.21	31.38	35.61	34.84	33.95
Qwen2.5-VL-72B-Instruct [2]	18.84	43.57	19.82	33.61	30.91	33.18	31.15	31.75	18.84	43.57	19.82	33.61	27.45	30.18	33.18	30.93
LLaVA-NeXT-Video-7B [72]	26.76	36.35	33.30	27.09	31.82	31.49	30.57	31.27	21.11	30.95	32.37	31.32	27.90	33.08	31.71	31.27
InternVL3-8B [59]	20.77	49.86	36.08	43.01	41.03	42.04	41.57	41.56	28.15	52.25	34.80	36.34	42.07	38.61	38.23	38.84
VideoLLaMA3-7B [68]	23.96	44.59	37.86	44.03	40.84	42.16	40.52	41.17	19.44	46.28	39.21	40.09	37.51	40.30	42.22	40.05
VideoLLaMA3-2B [68]	21.26	41.69	35.10	37.83	33.77	38.02	39.08	37.04	12.78	55.51	31.61	42.64	40.90	40.67	42.66	41.44
Qwen3-VL-8B-Instruct [1]	20.77	45.65	37.32	39.25	42.18	39.35	36.46	39.24	35.56	47.72	37.66	35.54	37.60	40.21	35.33	38.46
SFT baselines																
Qwen2.5-VL-7B [2]	46.67	76.30	53.02	72.53	75.04	63.82	61.77	66.68	80.00	65.77	54.10	64.74	63.40	63.92	63.68	63.67
Qwen2.5-VL-72B [2]	43.96	74.47	49.90	68.66	71.05	60.88	59.56	63.65	73.89	70.39	49.09	60.24	58.19	63.17	61.16	60.83
LLaVA-NeXT-Video-7B [72]	20.97	55.43	47.35	49.25	56.19	50.91	48.42	51.72	18.33	53.27	47.57	49.25	45.01	49.11	49.96	48.05
InternVL3-8B [59]	39.03	75.51	47.60	68.97	69.86	61.76	58.25	63.12	79.44	69.35	47.60	68.97	57.09	62.61	58.56	59.38
VideoLLaMA3-7B [68]	19.52	62.58	42.91	63.22	61.66	53.06	49.85	54.68	13.89	55.95	47.72	51.25	43.55	43.61	51.26	49.47
VideoLLaMA3-2B [68]	20.68	59.42	42.63	63.60	61.09	52.40	48.92	53.95	15.56	53.12	44.68	51.92	45.11	50.05	51.00	48.75
Qwen3-VL-8B-Instruct [1]	30.53	67.83	46.83	67.55	65.95	57.90	55.49	59.62	59.44	57.44	48.33	54.69	51.33	53.80	57.43	54.24
RAG Baselines (Best SFT model adopted)																
VideoRAG [26] (Qwen2.5-VL-7B-Instruct [2])	39.13	76.52	51.08	71.37	71.34	63.19	61.68	65.25	67.78	70.68	51.82	57.86	58.01	60.26	61.08	59.80
VideoRAG [26] (InternVL3-8B [59])	22.03	72.92	45.46	67.00	66.69	57.92	55.53	59.88	12.22	64.14	45.90	50.53	47.94	52.39	50.56	50.29
EgoRAG [63] (Qwen2.5-VL-7B-Instruct [2])	48.31	74.93	51.80	71.36	73.51	63.61	60.37	65.63	77.78	68.15	54.41	64.02	61.94	64.14	64.03	63.70
EgoRAG [63] (InternVL3-8B [59])	39.90	74.71	47.40	68.47	69.77	60.67	58.20	62.70	72.22	69.05	44.98	58.41	56.91	60.82	58.30	58.65
Keyframe Baselines (Best SFT model adopted)																
TSPO [57] (Qwen2.5-VL-7B-Instruct [2])	37.97	75.80	46.68	68.55	69.28	61.44	57.94	62.72	**86.11**	67.56	45.29	56.47	55.54	60.45	58.38	58.11
TSPO [57] (InternVL3-8B [59])	42.51	76.21	50.77	70.79	72.21	63.25	60.35	65.09	79.44	72.02	49.54	63.13	62.21	64.67	62.55	63.12
KTV [56] (Qwen2.5-VL-7B-Instruct [2])	19.71	52.32	41.68	57.11	50.56	49.63	47.33	49.13	21.67	53.12	43.47	42.03	38.98	45.55	45.79	43.46
KTV [56] (InternVL3-8B [59])	19.13	61.43	40.26	62.95	60.60	51.92	48.61	53.53	13.89	58.63	40.88	43.25	42.27	46.58	44.14	44.31
Our proposed approach																
SG-CoR (InternVL3-8B [59])	**57.78**	78.31	56.98	**74.96**	77.07	**68.07**	65.37	69.99	76.11	**73.81**	55.17	73.51	72.28	68.32	69.59	70.07
SG-CoR (Qwen3-VL-8B-Instruct [1])	57.20	78.14	58.38	74.54	**78.28**	67.32	65.31	70.11	79.44	72.77	**59.42**	73.18	72.37	**70.29**	69.50	70.70
SG-CoR (Qwen2.5-VL-7B-Instruct [2])	55.36	**78.72**	**59.36**	74.94	78.26	67.34	**66.60**	**70.55**	74.44	71.88	58.36	**74.63**	**72.74**	69.92	**69.85**	**70.82**

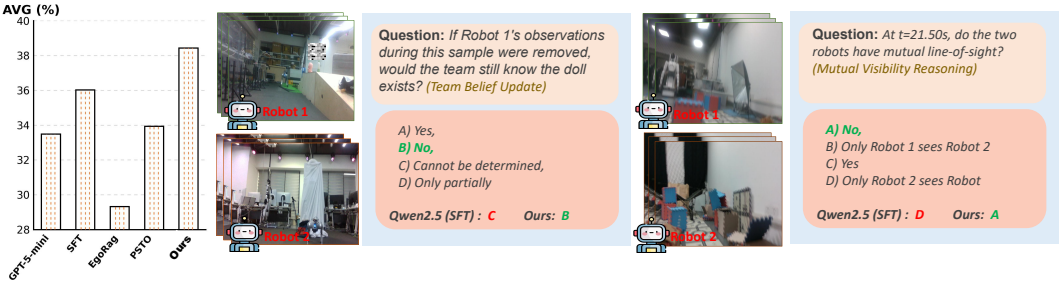


Figure 3: Performances of MLLMs on the real-world test set. Histograms on the left show the accuracy of different MLLMs. Samples on the right are qualitative samples.

Table 3: Ablation results on the Qwen2.5 backbone [2] for Habitat and iGibson test sets.

Method	Habitat [48] N=2	N=3	N=4	AVG	iGibson [32] N=2	N=3	N=4	AVG
w/o Spectral Energy Ranking in SE-MR^{2}FS	77.63	65.52	64.55	69.03	67.80	69.73	67.85	68.44
w/o Fourier Embed. in SGI-MRF	73.95	64.65	62.57	66.88	59.93	61.39	60.03	60.43
w/o SE-MR^{2}FS	72.53	63.75	61.00	65.58	61.94	65.98	59.51	62.40
w/o SGI-MRF	75.23	64.87	62.22	67.24	51.42	59.04	55.00	55.12
w/o GMIPSD	73.89	63.69	61.70	66.24	60.57	63.26	61.60	61.79
Ours	**78.26**	**67.34**	**66.60**	**70.55**	**72.74**	**69.92**	**69.85**	**70.82**

general-purpose MLLMs struggle with cooperative spatial reasoning from multi-robot egocentric videos. SFT substantially improves performance, for example, Qwen2.5-VL-7B-Instruct [2] increases from 33.51% to 66.68% on Habitat and from 33.95% to 63.67% on iGibson [32]. Accuracy differences across team sizes reflect variation in both team configuration and generated episodes, and therefore cannot be attributed solely to the number of robots. RAG and keyframe baselines such as EgoRAG [63] and TSPO [57] further improve evidence selection, but they mainly operate at the input level and do not explicitly model robot identity, inter-robot geometry, or cooperative view fusion. In contrast, **SG-CoR** achieves the best overall performance across both environments. The strongest **SG-CoR** variant reaches 70.55% on Habitat and 70.82% on iGibson, outperforming the strongest non-SG-CoR baselines by 3.87% and 7.12%, respectively. These gains indicate that the key challenge in **CoopSR** is not only selecting relevant frames, but also forming a consistent multi-robot scene representation. **SG-CoR** addresses this with three complementary biases: spectral sampling captures informative temporal changes for action and visibility reasoning; geometry- and

Table 4: Cross-team-size generalizability analysis on Habitat and iGibson when we train the model on $N = 2$ and $N = 3$ while we test on $N = 4$. Note that the averaged ACC is calculated on the averaged performance of all QAs.

Method	Habitat [48]					iGibson [32]				
	T=1	T=2	T=3	T=4	AVG	T=1	T=2	T=3	T=4	AVG
SFT (Qwen2.5-VL-7B-Instruct [2])	18.84	62.39	41.81	58.24	52.30	15.00	63.44	37.08	48.24	47.18
SFT (Qwen3-VL-8B-Instruct [1])	22.03	60.65	41.30	58.68	52.09	41.67	57.71	45.42	53.85	52.22
SFT (InternVL3-8B [59])	29.86	69.49	47.32	62.81	58.23	45.00	66.96	42.92	56.09	54.91
EgoRAG [63] (Qwen2.5-VL-7B-Instruct [2])	24.06	48.41	37.54	54.12	46.12	53.33	54.19	39.17	45.35	46.22
EgoRAG [63] (Qwen3-VL-8B-Instruct [1])	22.32	47.17	39.86	53.97	46.24	21.67	38.77	47.50	39.90	40.31
EgoRAG [63] (InternVL3-8B [59])	19.13	53.91	39.35	56.38	48.65	26.67	59.47	35.42	43.91	44.31
TSPO [57] (Qwen2.5-VL-7B-Instruct [2])	23.33	63.19	38.19	62.83	53.63	23.33	45.00	45.00	58.64	51.35
TSPO [57] (Qwen3-VL-8B-Instruct [1])	20.87	51.81	40.58	56.82	48.71	11.67	57.27	44.17	43.75	44.83
TSPO [57] (InternVL3-8B [59])	20.58	56.96	39.71	56.72	49.80	18.33	59.47	43.33	52.24	50.04
SG-CoR (Qwen2.5-VL-7B-Instruct [2])	50.14	70.94	51.81	68.79	63.56	71.67	66.96	51.25	71.15	66.20
SG-CoR (Qwen3-VL-8B-Instruct [1])	48.12	71.59	**54.57**	68.20	64.11	66.67	68.72	**56.67**	70.99	67.33
SG-CoR (InternVL3-8B [59])	**51.59**	**72.39**	53.48	**69.53**	**64.79**	**81.67**	**70.48**	55.42	**73.08**	**69.33**

motion-informed fusion aligns observations across changing viewpoints; and prompt distillation transfers trajectory-level geometric supervision to the student model without requiring ground-truth poses at test time. Together, these components improve cross-view consistency and team-level reasoning, leading to stronger gains on T4.

Analysis of ablation study. Table 3 ablates each component of **SG-CoR** with the Qwen2.5-VL-7B-Instruct backbone [2]. The full model achieves the best average accuracy on both Habitat [48] and iGibson [32], with 70.55% and 70.82%, respectively. Removing any component consistently degrades performance, showing that spectral sampling, Fourier temporal encoding, spectral-aware fusion, and geometry- and motion-informed prompt distillation are all beneficial. The largest drop occurs on iGibson [32] when SGI-MRF is removed, reducing accuracy from 70.82% to 55.12%, highlighting the importance of robust cross-view fusion in complex interactive scenes. Overall, the ablation confirms that these components are complementary and jointly improve multi-robot cooperative spatial reasoning.

Cross-team-size generalization. Table 4 evaluates generalization from smaller teams ($N = [2, 3]$) to an unseen larger team size ($N = 4$). This setting is challenging due to increased observation redundancy, viewpoint inconsistency, and more complex inter-agent relations. Across Habitat [48] and iGibson [32], **SG-CoR** achieves the best performance for all three backbones. The strongest **SG-CoR** variant reaches 64.79% on Habitat, improving over the best baseline by 6.56%, and 69.33% on iGibson [32], with a larger gain of 14.42%. These results show that **SG-CoR** generalizes more effectively than SFT, EgoRAG [63], and TSPO [57]. Beyond task adaptation, retrieval, or keyframe selection, **SG-CoR** combines dynamic spectral sampling with geometry- and motion-informed fusion and prompt distillation, explicitly modeling robot relations, co-visibility, and cross-view alignment. These cooperative inductive biases enable stronger transfer to the unseen $N = 4$ setting.

Real-world generalizability. Figure 3 shows that **SG-CoR** achieves the best real-world accuracy, outperforming GPT-5 mini [55], SFT, EgoRAG [63], and TSPO [57]. Qualitative examples further show that **SG-CoR** handles team belief updates and mutual visibility reasoning more reliably than SFT, indicating stronger generalizability to viewpoint changes, partial observations, and multi-robot spatial relations in real-world situations.

6 Conclusion

We introduced **CoopSR**, a benchmark for multi-robot cooperative egocentric spatial reasoning, requiring models to integrate synchronized views for spatial, temporal, visibility, and coordination reasoning. We further proposed **SG-CoR**, which combines spectral frame sampling, geometry- and motion-informed fusion, and prompt distillation to build coherent team-level representations without requiring privileged pose information at test time. Experiments on Habitat, iGibson, cross-team-size evaluation, and real-world robot data show that **SG-CoR** consistently outperforms strong MLLM, SFT, RAG, and keyframe baselines, highlighting the importance of physically grounded cooperative reasoning for embodied AI.

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[3] Sven Bambach, Stefan Lee, David J. Crandall, and Chen Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *ICCV*, 2015.

[4] Siddhant Bansal, Chetan Arora, and C.V. Jawahar. My view is the best view: Procedure learning from egocentric videos. In *ECCV*, 2022.

[5] Andy Bonnetto, Haozhe Qi, Franklin Leong, Matea Tashkovska, Mahdi Rad, Solaiman Shokur, Friedhelm Hummel, Silvestro Micera, Marc Pollefeys, and Alexander Mathis. EPFL-Smart-Kitchen: An ego-exo multi-modal dataset for challenging action and motion understanding in video-language models. In *NeurIPS*, 2025.

[6] Qi Chai, Zhang Zheng, Junlong Ren, Deheng Ye, Zichuan Lin, and Hao Wang. CausalMACE: Causality empowered multi-agents in minecraft cooperative tasks. In *EMNLP (Findings)*, 2025.

[7] Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. HourVideo: 1-hour video-language understanding. In *NeurIPS*, 2024.

[8] Matthew Chang, Gunjan Chhablani, Alexander Clegg, Mikael Dallaire Cote, Ruta Desai, Michal Hlavac, Vladimir Karashchuk, Jacob Krantz, Roozbeh Mottaghi, Priyam Parashar, Siddharth Patki, Ishita Prasad, Xavier Puig, Akshara Rai, Ram Ramrakhya, Daniel Tran, Joanne Truong, John M. Turner, Eric Undersander, and Tsung-Yen Yang. PARTNR: A benchmark for planning and reasoning in embodied multi-agent tasks. *arXiv preprint arXiv:2411.00081*, 2024.

[9] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. CrossViT: Cross-attention multi-scale vision transformer for image classification. In *ICCV*, 2021.

[10] Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. EgoPlan-Bench: Benchmarking multimodal large language models for human-level planning. *International Journal of Computer Vision*, 2026.

[11] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024.

[12] Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. EgoThink: Evaluating first-person perspective thinking capability of vision-language models. In *CVPR*, 2024.

[13] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The EPIC-KITCHENS dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[14] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. EPIC-KITCHENS VISOR benchmark: Video segmentations and object relations. In *NeurIPS*, 2022.

[15] Fernando De la Torre, Jessica Hodgins, Adam Bargteil, Xavier Martin, Justin Macey, Alex Collado, and Pep Beltran. Guide to the carnegie mellon university multimodal activity (CMU-MMAC) database. 2009.

[16] Chenyou Fan. EgoVQA-an egocentric video question answering benchmark dataset. In *ICCVW*, 2019.

[17] Alircza Fathi, Jessica K. Hodgins, and James M. Rehg. Social interactions: A first-person perspective. In *CVPR*, 2012.

[18] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022.

[19] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, et al. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In *CVPR*, 2024.

[20] Xudong Guo, Kaixuan Huang, Jiale Liu, Wenhui Fan, Natalia Vélez, Qingyun Wu, Huazheng Wang, Thomas L. Griffiths, and Mengdi Wang. Embodied LLM agents learn to cooperate in organized teams. *IEEE Transactions on Computational Social Systems*, 2026.

[21] 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. *arXiv preprint arXiv:2507.18342*, 2025.

[22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 2022.

[23] Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, 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 *CVPR*, 2024.

[24] Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2.5-Coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

[25] Unnat Jain, Luca Weihs, Eric Kolve, Ali Farhadi, Svetlana Lazebnik, Aniruddha Kembhavi, and Alexander Schwing. A cordial sync: Going beyond marginal policies for multi-agent embodied tasks. In *ECCV*, 2020.

[26] Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. VideoRAG: Retrieval-augmented generation over video corpus. In *ACL (Findings)*, 2025.

[27] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. EgoTaskQA: Understanding human tasks in egocentric videos. In *NeurIPS*, 2022.

[28] Minjoon Jung, Junbin Xiao, Junghyun Kim, Byoung-Tak Zhang, and Angela Yao. EgoExo-Con: Exploring view-invariant video temporal understanding. *arXiv preprint arXiv:2510.26113*, 2025.

[29] Kangsan Kim, Yanlai Yang, Suji Kim, Woongyeong Yeo, Youngwan Lee, Mengye Ren, and Sung Ju Hwang. MA-EgoQA: Question answering over egocentric videos from multiple embodied agents. *arXiv preprint arXiv:2603.09827*, 2026.

[30] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012.

[31] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench: Benchmarking multimodal large language models. In *CVPR*, 2024.

[32] Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Elliott Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, Andrey Kurenkov, C. Karen Liu, Hyowon Gweon, Jiajun Wu, Li Fei-Fei, and Silvio Savarese. iGibson 2.0: Object-centric simulation for robot learning of everyday household tasks. In *CoRL*, 2021.

[33] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *EMNLP*, 2024.

[34] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, Chengfei Cai, Hongfa Wang, Dima Damen, Bernard Ghanem, Wei Liu, and Mike Zheng Shou. Egocentric video-language pretraining. In *NeurIPS*, 2022.

[35] Benlin Liu, Yuhao Dong, Yiqin Wang, Zixian Ma, Yansong Tang, Luming Tang, Yongming Rao, Wei-Chiu Ma, and Ranjay Krishna. Coarse correspondences boost spatial-temporal reasoning in multimodal language model. In *CVPR*, 2025.

[36] Kehui Liu, Zixin Tang, Dong Wang, Zhigang Wang, Xuelong Li, and Bin Zhao. COHERENT: Collaboration of heterogeneous multi-robot system with large language models. In *ICRA*, 2025.

[37] Youzhi Liu, Li Gao, Liu Liu, Mingyang Lv, and Yang Cai. CoMaTrack: Competitive multi-agent game-theoretic tracking with vision-language-action models. *arXiv preprint arXiv:2603.22846*, 2026.

[38] Zhongyuang Liu, Min He, Shaonan Yu, Xinhang Xu, Muqing Cao, Jianping Li, Jianfei Yang, and Lihua Xie. OmniVLN: Omnidirectional 3D perception and token-efficient LLM reasoning for visual-language navigation across air and ground platforms. *arXiv preprint arXiv:2603.17351*, 2026.

[39] Qian Long, Zhi Li, Ran Gong, Ying Nian Wu, Demetri Terzopoulos, and Xiaofeng Gao. TeamCraft: A benchmark for multi-modal multi-agent systems in minecraft. *arXiv preprint arXiv:2412.05255*, 2024.

[40] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[41] Yongdong Luo, Xiawu Zheng, Xiao Yang, Guilin Li, Haojia Lin, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. Video-RAG: Visually-aligned retrieval-augmented long video comprehension. In *NeurIPS*, 2025.

[42] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *NeurIPS*, 2023.

[43] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric RGB-D sensor. In *ICCV*, 2017.

[44] OpenAI. GPT-4o system card, 2024.

[45] Andrea Palazzi, Davide Abati, Simone Calderara, Francesco Solera, and Rita Cucchiara. Predicting the driver's focus of attention: The DR(eye)VE project. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[46] Junhao Pan, Zehua Yuan, Xiaofan Zhang, and Deming Chen. YouHome system and dataset: Making your home know you better. In *iSES*, 2022.

[47] Chiara Plizzari, Mirco Planamente, Gabriele Goletto, Marco Cannici, Emanuele Gusso, Matteo Matteucci, and Barbara Caputo. E^2(GO)MOTION: Motion augmented event stream for egocentric action recognition. In *CVPR*, 2022.

[48] Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander Clegg, Michal Hlavac, So Yeon Min, Vladimir Vondrus, Théophile Gervet, Vincent-Pierre Berges, John M. Turner, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Singh Chaplot, Unnat Jain, Dhruv Batra, Akshara Rai, and Roozbeh Mottaghi. Habitat 3.0: A co-habitat for humans, avatars and robots. *arXiv preprint arXiv:2310.13724*, 2023.

[49] Yiran Qin, Li Kang, Xiufeng Song, Zhenfei Yin, Xiaohong Liu, Xihui Liu, Ruimao Zhang, and Lei Bai. RoboFactory: Exploring embodied agent collaboration with compositional constraints. In *ICCV*, 2025.

[50] Heqian Qiu, Zhaofeng Shi, Lanxiao Wang, Huiyu Xiong, Xiang Li, and Hongliang Li. EgoMe: A new dataset and challenge for following me via egocentric view in real world. *arXiv preprint arXiv:2501.19061*, 2025.

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[52] Michael S Ryoo and Larry Matthies. First-person activity recognition: What are they doing to me? In *CVPR*, 2013.

[53] Georgy Savva, Oscar Michel, Daohan Lu, Suppakit Waiwitlikhit, Timothy Meehan, Dhairya Mishra, Srivats Poddar, Jack Lu, and Saining Xie. Solaris: Building a multiplayer video world model in minecraft. *arXiv preprint arXiv:2602.22208*, 2026.

[54] Tim J. Schoonbeek, Tim Houben, Hans Onvlee, Peter H. N. de With, and Fons van der Sommen. IndustReal: A dataset for procedure step recognition handling execution errors in egocentric videos in an industrial-like setting. In *WACV*, 2024.

[55] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[56] Baiyang Song, Jun Peng, Yuxin Zhang, Guangyao Chen, Feidiao Yang, and Jianyuan Guo. KTV: Keyframes and key tokens selection for efficient training-free video LLMs. *arXiv preprint arXiv:2602.03615*, 2026.

[57] Canhui Tang, Zifan Han, Hongbo Sun, Sanping Zhou, Xuchong Zhang, Xin Wei, Ye Yuan, Huayu Zhang, Jinglin Xu, and Hao Sun. TSPO: Temporal sampling policy optimization for long-form video language understanding. In *AAAI*, 2026.

[58] Pavel Tokmakov, Jie Li, and Adrien Gaidon. Breaking the "object" in video object segmentation. In *CVPR*, 2023.

[59] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[60] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. HoloAssist: an egocentric human interaction dataset for interactive AI assistants in the real world. In *ICCV*, 2023.

[61] Yaxuan Wang, Yifan Xiang, Ke Li, Xun Zhang, BoWen Ye, Zhuochen Fan, Fei Wei, and Tong Yang. Can a robot walk the robotic dog: Triple-zero collaborative navigation for heterogeneous multi-agent systems. *arXiv preprint arXiv:2603.21723*, 2026.

[62] Haoyu Wu, Jiwen Yu, Yingtian Zou, and Xihui Liu. MultiWorld: Scalable multi-agent multi-view video world models. *arXiv preprint arXiv:2604.18564*, 2026.

[63] Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Bo Li, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, and Ziwei Liu. EgoLife: Towards egocentric life assistant. In *CVPR*, 2025.

[64] Hanrong Ye, Haotian Zhang, Erik A. Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Bowen Zhang, Haoxuan You, Dan Xu, Zhe Gan, Jiasen Lu, and Yinfei Yang. MM-Ego: Towards building egocentric multimodal LLMs. *arXiv preprint arXiv:2410.07177*, 2024.

[65] Bangguo Yu, Hamidreza Kasaei, and Ming Cao. Co-NavGPT: Multi-robot cooperative visual semantic navigation using vision language models. *arXiv preprint arXiv:2310.07937*, 2023.

[66] Chao Yu, Xinyi Yang, Jiaxuan Gao, Jiayu Chen, Yunfei Li, Jijia Liu, Yunfei Xiang, Ruixin Huang, Huazhong Yang, Yi Wu, and Yu Wang. Asynchronous multi-agent reinforcement learning for efficient real-time multi-robot cooperative exploration. In *AAMAS*, 2023.

[67] Jirong Zha, Yuxuan Fan, Tianyu Zhang, Geng Chen, Yingfeng Chen, Chen Gao, and Xinlei Chen. AirCopBench: A benchmark for multi-drone collaborative embodied perception and reasoning. In *AAAI*, 2026.

[68] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. VideoLLaMA 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

[69] Deheng Zhang, Yuqian Fu, Runyi Yang, Yang Miao, Tianwen Qian, Xu Zheng, Guolei Sun, Ajad Chhatkuli, Xuanjing Huang, Yu-Gang Jiang, Luc Van Gool, and Danda Pani Paudel. EgoNight: Towards egocentric vision understanding at night with a challenging benchmark. In *ICLR*, 2026.

[70] Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B. Tenenbaum, Tianmin Shu, and Chuang Gan. Building cooperative embodied agents modularly with large language models. In *ICLR*, 2024.

[71] Hongxin Zhang, Zeyuan Wang, Qiushi Lyu, Zheyuan Zhang, Sunli Chen, Tianmin Shu, Behzad Dariush, Kwonjoon Lee, Yilun Du, and Chuang Gan. COMBO: Compositional world models for embodied multi-agent cooperation. In *ICLR*, 2025.

[72] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. LLaVA-NeXT: A strong zero-shot video understanding model, April 2024.

[73] Zijie Zhao, Honglei Guo, Shengqian Chen, Kaixuan Xu, Bo Jiang, Yuanheng Zhu, and Dongbin Zhao. Empowering multi-robot cooperation via sequential world models. *arXiv preprint arXiv:2509.13095*, 2025.

[74] Sunyao Zhou, Yunzi Wu, Tianhang Wang, Xinhai Li, Guang Chen, Lizheng Liu, Chenjia Bai, and Xuelong Li. DeCoNav: Dialog enhanced long-horizon collaborative vision-language navigation. *arXiv preprint arXiv:2604.12486*, 2026.

[75] Bingwen Zhu, Yuqian Fu, Qiaole Dong, Guolei Sun, Tianwen Qian, Yuzheng Wu, Danda Pani Paudel, Xiangyang Xue, and Yanwei Fu. EgoSound: Benchmarking sound understanding in egocentric videos. In *CVPR*, 2026.

Acknowledgments and Disclosure of Funding

The project is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1574 – 471687386. This work was performed on the HoreKa supercomputer funded by the Ministry of Science, Research and the Arts Baden-Württemberg and by the Federal Ministry of Education and Research. The authors also acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG. This project is also supported in part by the National Natural Science Foundation of China under Grant No. 62473139, No. 62673195, and No. 62388101, in part by the Hunan Provincial Research and Development Project (Grant No. 2025QK3019), and in part by the State Key Laboratory of Autonomous Intelligent Unmanned Systems (the opening project number ZZKF2025-2-10). This research was partially funded by the Ministry of Education and Science of Bulgaria (support for INSAIT, part of the Bulgarian National Roadmap for Research Infrastructure).

A Appendices

A.1 Societal Impact and Limitations

This work advances cooperative spatial reasoning for multi-robot teams that jointly interpret synchronized egocentric observations. By enabling robots to integrate partial viewpoints into a shared

scene understanding, **CoopSR** may support important applications in search and rescue, assistive robotics, smart homes, warehouse automation, infrastructure inspection, and navigation in hazardous or unknown environments. In these settings, stronger team-level situational awareness can help robots cover larger areas, resolve occlusions, coordinate tasks more effectively, and reduce human exposure to risk.

At the same time, multi-robot egocentric perception introduces privacy and safety challenges. Robots operating in homes, workplaces, hospitals, or public spaces may capture sensitive visual information, so deployment should follow data minimization, consent, anonymization, secure storage, and access-control practices. In addition, higher benchmark accuracy should not be interpreted as a guarantee of safe autonomous behavior. False spatial reasoning, incorrect visibility estimation, or wrong task assignment could lead to unsafe navigation, collisions, failed coordination, or inappropriate actions in physical environments. Therefore, real-world deployment should include human oversight, uncertainty estimation, fail-safe mechanisms, and extensive validation under diverse operating conditions.

Our dataset is primarily simulation-based, with a controlled real-world quadruped robot test set. While this enables scalable and systematic evaluation, it may not fully capture open-world variation such as clutter, lighting changes, sensor noise, social context, larger robot teams, or unseen embodiments. Cooperative robot reasoning also has dual-use potential: it can improve rescue, accessibility, and industrial safety, but could also be misused for surveillance or intrusive monitoring. Future work should therefore evaluate robustness, privacy preservation, fairness, and human-centered safety alongside accuracy.

Limitation Discussion. Although **CoopSR** covers diverse simulated environments, its real-world evaluation remains limited in scene diversity, robot types, and team sizes. The benchmark deliberately controls embodiment to isolate cooperative spatial reasoning, while the Go2/Go2-W team introduces only limited within-family platform variation. **SG-CoR** processes per-agent RGB streams without requiring robot poses at inference, but this architectural flexibility does not establish robustness to arbitrary embodiments or sensing suites.

Future work should evaluate larger heterogeneous teams, including aerial–ground teams and unseen embodiments, while systematically varying camera intrinsics, field of view, sensor quality, robot height, and locomotion type. Extending beyond RGB observations would additionally require modality-specific encoders. Broader evaluation should also encompass open-ended tasks and safety-critical scenarios with richer physical interactions.

A.2 Details of Simulator and Dataset Splits

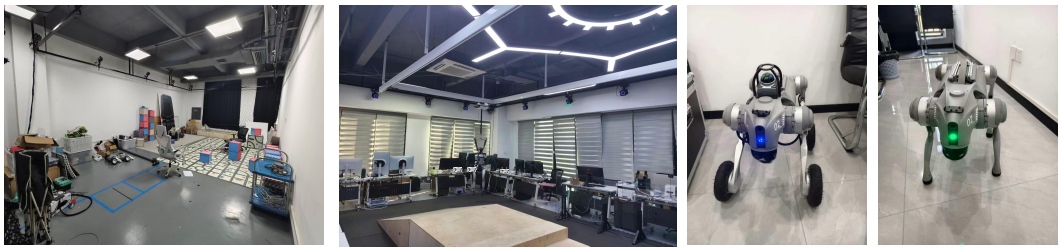


Figure 4: Overview of data collection devices. We make use of two Unitree quadruped robots and two different motion capture systems from FZMotion and NOKOV. We use daily life objects to set up 13 different scenarios in these two data collection rooms, which can provide ground truth pose data for the robots.

Habitat scenes. Habitat [48] is a high-performance embodied AI simulation platform that supports photorealistic indoor navigation, egocentric RGB-D observation, and scalable agent-environment interaction. Robots are deployed in 154 procedurally generated indoor floor plans. Each scene is explored by $N \in \{2, 3, 4\}$ robots executing a coverage exploration policy. Synchronised RGB video, semantic segmentation mask for each frame, depth video, odometry commands, and ground-truth

poses are recorded. The trajectory log for each exploration is stored together with the co-located video files, containing per-timestep columns: position, heading, forward velocity, and angular velocity.

iGibson scenes. IGibson [32] is a realistic interactive simulation platform for embodied AI, supporting indoor navigation, geometry- and motion-informed interaction, and egocentric RGB-D perception in diverse household environments. 14 interactive scenes are adopted for the data collection. During the exploration, we also adopt $N \in \{2, 3, 4\}$ for the data collection.

Train / val / test splits. For each configuration, scenes are divided into disjoint training, validation, and test splits, with $119/12/23$ scenes for Habitat settings and $8/2/4$ scenes for iGibson. For the real-world quadruped robot set, all collected scenes are used for testing. No scene is shared across splits, thereby preventing scene-level data leakage.

Episode generation across team sizes. The scene name and exploration index initialize each simulator run, after which robot starting positions, waypoints, and motions are generated jointly for the specified team size. Habitat computes rendezvous and tether targets from the mean position of all active robots, while iGibson computes NEAR-phase waypoints from the full-team centroid. Consequently, changing N changes the trajectories and observations even for the same scene and exploration index. A future paired observation evaluation could hold an $N = 4$ episode fixed and mask predefined robot streams to evaluate $N = \{2, 3\}$, isolating the effect of available views while preserving the underlying trajectories.

A.3 Dataset Annotation Details, Diversity and Statistical Analysis

A.3.1 Human Checking Details

To ensure the reliability of **CoopSR**, we conduct a multi-stage human verification process. Four human annotators independently review the generated QA pairs using synchronized multi-robot egocentric videos, simulator metadata, and the proposed answer choices. Annotators verify whether each question is visually grounded, whether the answer is uniquely determined, and whether the question genuinely requires cooperative reasoning across multiple robot viewpoints rather than single-view recognition or language priors.

For each QA pair, annotators either accept it, correct it, or discard it. Corrections include resolving robot identity confusion, inaccurate spatial relations, ambiguous object references, incorrect visibility states, and temporal boundary inconsistencies. QA pairs are discarded if the target object is not sufficiently observable, if multiple answer choices are plausible, if the answer can be inferred from textual priors alone, or if a cross-robot level question does not require cross-robot evidence despite belonging to a cooperative reasoning category. To further improve benchmark quality, QA pairs marked ambiguous by at least two annotators are removed.

A.3.2 Dataset Diversity and Statistical Analysis

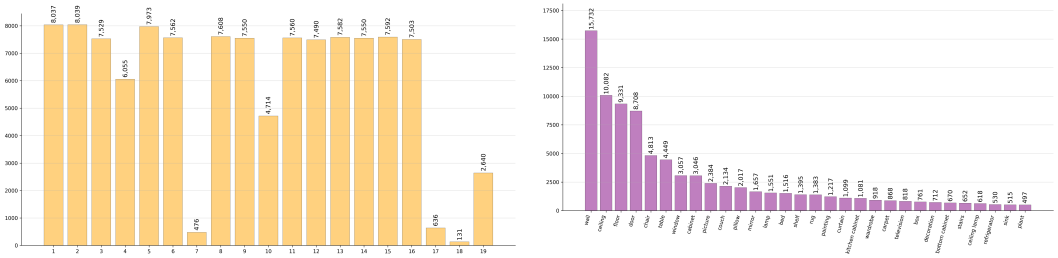
We analyze the linguistic and semantic diversity of **CoopSR** in Figures 5 and 6. Figure 5 shows word clouds for the Habitat [48], iGibson [32], and real-world lab subsets. Across domains, frequent terms such as "timestep", "observations", "robot", "object", "team", and "between" indicate a shared focus on temporally grounded multi-robot egocentric reasoning. Habitat emphasizes spatial relations with terms such as "left", "right", "distance", and "closest", whereas iGibson [32] highlights interaction-rich household scenarios through words such as "chair", "cabinet", "table", "action", and "pushing". The real-world subset contains prominent terms such as "pose", "nearest", "farthest", "field", "region", and "depth", reflecting visibility-, distance-, and pose-aware reasoning.

Overall, Figure 5 suggests that **CoopSR** spans diverse simulation and real-world language patterns while consistently targeting cooperative spatial, temporal, visibility, and action reasoning. Figure 5 provides quantitative statistics of the QA and object distributions. Figure 6a reports the number of QA samples for the 19 QA types. Most QA categories contain approximately $7K$–$8K$ samples, indicating broad and relatively balanced coverage across the benchmark taxonomy. A few categories contain fewer samples because they correspond to more constrained events or higher-level cooperative reasoning cases that require specific scene configurations, such as rare interaction events, team-level belief updates, or task-assignment conditions. This distribution ensures that the benchmark contains sufficient supervision for common spatial and temporal reasoning tasks while still including challenging long-tail cooperative reasoning cases.



(a) Habitat set [48] (b) iGibson set [32] (c) Real-world set

Figure 5: An overview of the word clouds of the QAs from (a) Habitat, (b) iGibson, and (c) Real-world lab.



(a) QA statistics. (b) Object statistics from metadata.

Figure 6: An overview of the statistics of (a) the number of QAs and (b) the number of objects.

Figure 6b summarizes the object-frequency distribution extracted from metadata. The object distribution is naturally long-tailed: structural and frequently visible categories such as walls, chairs, floors, doors, and cabinets appear most often, while smaller or scene-specific objects such as pillows, lamps, mirrors, sinks, flowers, televisions, and decorative objects occur less frequently. This long-tail distribution reflects realistic indoor environments, where common background structures dominate visual observations while many task-relevant objects appear sparsely. Such diversity is important for evaluating whether MLLMs can reason beyond frequent object categories and generalize to less common objects under partial, egocentric, and multi-view observations.

Together, Figures 5 and 6 demonstrate that **CoopSR** provides both linguistic diversity and semantic richness. The benchmark is not limited to simple object recognition or single-view spatial questions; instead, it contains diverse question formulations, multiple reasoning types, varied object categories, and both simulated and real-world domains. These statistics support the effectiveness of **CoopSR** as a comprehensive benchmark for evaluating cooperative multi-robot egocentric spatial reasoning.

A.4 More Ablation Analysis of our SG-CoR Approach on Hyperparameters

We further analyze the sensitivity of **SG-CoR** to three important hyperparameters: the distillation loss weight in GMIPSD, the number of learnable prompts, and the number of dense frames retained during CLIP-based similarity pre-selection. The results are reported on both Habitat [48] and iGibson [32] across different robot-team sizes. Table 5a compares different fusion strategies in SGI-MRF. Overall, the proposed fusion strategy consistently achieves the best performance on both Habitat and iGibson, with average accuracies of 70.55% and 70.82%, respectively. Compared with simple concatenation, addition, and multiplication, our method provides more effective multi-view feature integration, indicating that directly combining view features is insufficient to fully capture complementary spatial cues. Self-attention performs better than these simple fusion baselines, suggesting that adaptive view-wise feature interaction is beneficial. However, our method further improves over self-attention by 1.56% on Habitat [48] and 4.59% on iGibson [32] in terms of average accuracy, demonstrating its stronger capability in modeling cross-view relationships and selecting informative features across different numbers of observed views.

First, we vary the weight of the prompt-space distillation loss, λ_d, from 0.1 to 0.7 in Table 5b. The results show that GMIPSD is effective across a broad range of weights. On Habitat [48], the average accuracy remains stable, ranging from 69.84% to 70.55%, indicating that the model is not overly sensitive to the exact distillation weight. On iGibson [32], increasing λ_d generally

Table 5: Ablation studies of SGI-MRF and GMIPSD.

(a) Ablation study of the fusion strategy used in SGI-MRF.

Method	Habitat [48]				iGibson [32]			
	N=2	N=3	N=4	AVG	N=2	N=3	N=4	AVG
Concatenation	72.17	60.55	56.66	62.89	60.66	62.23	62.12	61.67
Addition	65.61	58.61	54.56	59.43	46.75	51.64	51.95	50.14
Multiplication	69.24	59.87	56.76	61.77	51.51	57.73	55.08	54.76
SelfAttention	75.65	66.80	65.00	68.99	64.78	67.48	66.46	66.23
Ours	**78.26**	**67.34**	**66.60**	**70.55**	**72.74**	**69.92**	**69.85**	**70.82**

(b) Ablation of the loss weight for L_{distill} ranging from $[0.1, 0.7]$.

L_{distill} in GMIPSD	Habitat [48]				iGibson [32]			
	N=2	N=3	N=4	AVG	N=2	N=3	N=4	AVG
0.1	77.63	67.62	64.86	69.84	68.25	68.79	67.33	68.11
0.3	**78.26**	67.34	66.60	**70.55**	**72.74**	69.92	69.85	70.82
0.5	72.19	**71.13**	**70.81**	70.54	72.19	71.13	**70.81**	71.37
0.7	78.12	66.84	65.76	70.05	72.46	**72.63**	69.50	**71.49**

Table 6: Ablation studies of dense frame selection and learnable prompts.

(a) Ablation for dense frame selection during the CLIP-based similarity pre-selection.

#Dense Frames	Habitat [48]				iGibson [32]			
	N=2	N=3	N=4	AVG	N=2	N=3	N=4	AVG
16	76.91	**67.66**	65.88	69.98	69.26	69.17	67.94	68.77
32	**78.26**	67.34	**66.60**	**70.55**	**72.74**	**69.92**	**69.85**	**70.82**
64	77.48	66.62	64.46	69.32	67.52	69.17	67.68	68.11

(b) Ablation of the number of learnable prompts ranging from $[4, 10]$.

#Prompts in GMIPSD	Habitat [48]				iGibson [32]			
	N=2	N=3	N=4	AVG	N=2	N=3	N=4	AVG
4	77.12	67.70	**66.69**	70.34	69.99	69.82	67.33	69.01
6	77.20	67.66	64.88	69.73	69.99	69.82	67.16	68.95
8	**78.26**	67.34	66.60	**70.55**	**72.74**	**69.92**	**69.85**	**70.82**
10	77.65	**68.05**	65.21	70.11	68.07	69.82	68.55	68.80

Table 7: Comparison to text-only baseline on **CoopSR** benchmark.

Datasets / **Models**	**Habitat** [48]				**iGibson** [32]			
	N=2	**N=3**	**N=4**	**AVG**	**N=2**	**N=3**	**N=4**	**AVG**
Text-Only Baseline	23.51	27.27	29.97	29.47	26.58	30.31	31.30	26.97
SFT (Qwen2.5-VL-7B-Instruct [2])	75.04	63.82	61.77	66.68	63.40	63.92	63.68	63.67
SG-CoR (Qwen2.5-VL-7B-Instruct [2])	**78.26**	**67.34**	**66.60**	**70.55**	**72.74**	**69.92**	**69.85**	**70.82**

improves performance, with the average accuracy increasing from 68.11% at $\lambda_d = 0.1$ to 71.49% at $\lambda_d = 0.7$. This suggests that stronger geometry- and motion-informed prompt supervision is particularly beneficial in iGibson [32], where scenes contain more complex interactive dynamics. However, overly large weights may slightly reduce performance for some team-size settings, implying that the distillation objective should complement, rather than dominate, the task supervision.

Second, we study the number of learnable prompts in GMIPSD as shown in Table 6b. The performance remains relatively stable when using 4 to 10 prompts, demonstrating the robustness of the prompt-space distillation design. Using 8 prompts achieves a strong balance across both environments, reaching 69.99% average accuracy on Habitat [48] and 70.07% on iGibson [32]. Increasing the number of prompts to 10 slightly improves Habitat performance to 70.11%, but decreases iGibson [32] accuracy to 68.80%. This indicates that a larger prompt capacity does not necessarily lead to better generalization, likely because excessive prompts introduce redundant latent factors or overfit to environment-specific patterns. Therefore, a moderate number of prompts provides sufficient capacity to encode distilled spatial priors while maintaining stable generalization.

Third, we evaluate the number of dense frames used during CLIP-based similarity pre-selection, as shown in Table 6a. Increasing the dense-frame budget from 16 to 32 improves the average accuracy from 69.98% to 70.55% on Habitat [48] and from 68.77% to 70.82% on iGibson [32]. This confirms that retaining more candidate frames before spectral refinement helps preserve useful temporal and semantic evidence. However, further increasing the budget to 64 does not yield additional gains, leading to 69.99% on Habitat and 70.07% on iGibson [32]. This suggests that a moderate candidate set is sufficient for capturing informative moments, while too many dense frames may introduce redundant or noisy visual evidence.

Overall, these ablations show that hyperparameter sensitivity of **SG-CoR** is acceptable. GMIPSD benefits from a properly weighted distillation objective and a moderate number of learnable prompts, while the frame sampler performs best when the CLIP-based candidate set balances evidence coverage and redundancy reduction. These findings further support the design of **SG-CoR**, where geometry- and motion-informed prompt distillation and spectral frame selection jointly improve cooperative multi-robot spatial reasoning.

Table 8: Comparison of our approach and the counterpart under robot Dropout setting.

Datasets	Habitat [48]					iGibson [32]				
Models	**T=1**	**T=2**	**T=3**	**T=4**	**AVG**	**T=1**	**T=2**	**T=3**	**T=4**	**AVG**
Robot Dropout on SG-CoR (Qwen2.5-VL-7B-Instruct [2])	31.69	58.29	44.10	59.74	53.29	48.89	52.38	43.31	48.64	48.35
SG-CoR (Qwen2.5-VL-7B-Instruct [2])	**55.36**	**78.72**	**59.36**	**74.94**	**70.55**	**74.44**	**71.88**	**58.36**	**74.63**	**70.82**

A.5 More Qualitative Analysis

Figure 7 to Figure 12 compare **SG-CoR** with the SFT (Qwen2.5-VL-7B [2]) baseline on representative **CoopSR** examples. **SG-CoR** is more accurate on tasks requiring cross-robot view integration, temporal event localization, mutual visibility reasoning, and teamwise object association. The SFT baseline often confuses robot identities, action phases, or visually similar objects across views, suggesting that standard fine-tuning mainly adapts the backbone to the QA format but does not provide explicit cooperative inductive biases.

These examples highlight the novelty of **SG-CoR**: rather than treating multi-robot videos as concatenated visual tokens, it introduces dynamics-aware frame sampling to retain informative temporal evidence, geometry- and motion-informed fusion to align observations under changing egocentric viewpoints, and prompt-space distillation to transfer pose-level geometric priors without requiring poses at test time. As a result, **SG-CoR** can build a more consistent team-level scene representation, enabling robust reasoning over partial observations, inter-robot relations, and dynamic cooperative events.

A.6 Comparison with Text-Only Baseline

We include a text-only baseline to evaluate whether **CoopSR** can be solved using linguistic priors or dataset biases without visual observations. As shown in Table 7, the text-only baseline obtains much lower accuracy than vision-based models, with only 29.47% on Habitat and 26.97% on iGibson, compared with 66.68% and 63.67% for the SFT baseline. This large gap demonstrates that **CoopSR** requires grounding in synchronized multi-robot egocentric videos, while the further gains of **SG-CoR** show the benefit of explicit cooperative spatial reasoning.

A.7 Importance of Multi-Robot Cooperative Reasoning and Experiments of Missing Robot

To verify that **CoopSR** genuinely requires cooperative reasoning across multiple robots rather than relying on a single informative viewpoint, we conduct a robot-dropout analysis, where one robot stream is randomly removed during inference. Results are shown in Table 8. Removing one robot consistently degrades performance across all reasoning tiers and datasets. On Habitat, the average accuracy drops from 70.55% to 53.29%, while on iGibson [32] it decreases from 70.82% to 48.35%.

The degradation is especially clear on higher-level cooperative reasoning tasks, where models must integrate complementary observations to infer scene-level structure, mutual visibility, temporal interactions, and team belief updates. These results show that **CoopSR** cannot be reliably solved from a reduced set of robot views alone. Missing one robot removes critical evidence for resolving occlusions, cross-view object grounding, and inter-robot spatial relations. The substantial gap between Robot Dropout and full **SG-CoR** confirms that the benchmark evaluates genuine multi-robot cooperative spatial reasoning rather than independent single-robot perception.

A.8 Additional Communication Baselines and Failure Analysis

Decentralized language-communication baselines. We evaluate two additional baselines on the Habitat test set, with all methods using the Qwen2.5-VL-7B backbone. These baselines process each robot's video independently before combining information through language or voting. In *Structured Report Aggregator*, each robot converts its observations into a robot-indexed textual report, and an LLM aggregates the reports to answer the question. In *Confidence-Weighted Vote*, each robot independently predicts an answer and confidence score, and the final answer is selected by confidence-weighted voting.

Quantitative comparison. As shown in Table 9, Structured Report Aggregator outperforms both Confidence-Weighted Vote and the centralized zero-shot baseline across all reported reasoning tiers

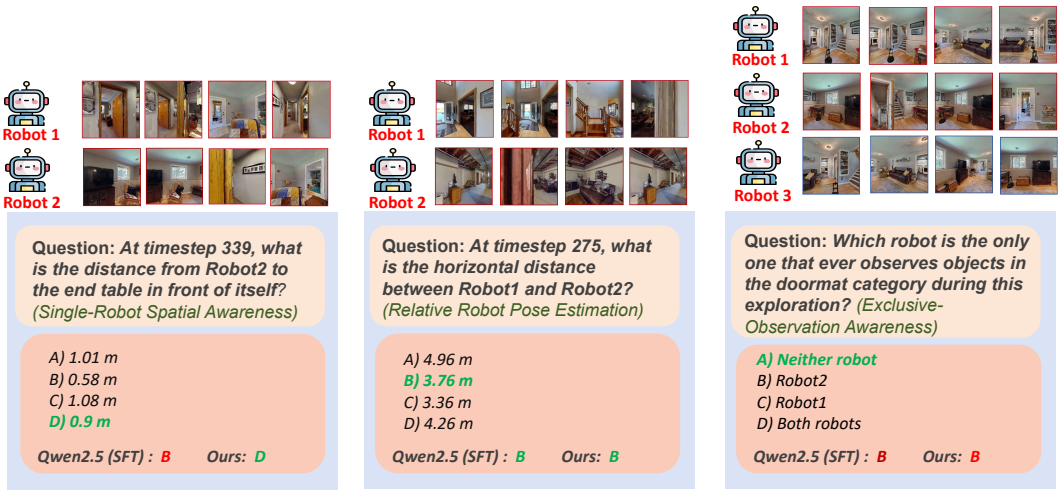


Figure 7: Overview of qualitative results of our approach and SFT Qwen2.5-VL-7B [24].

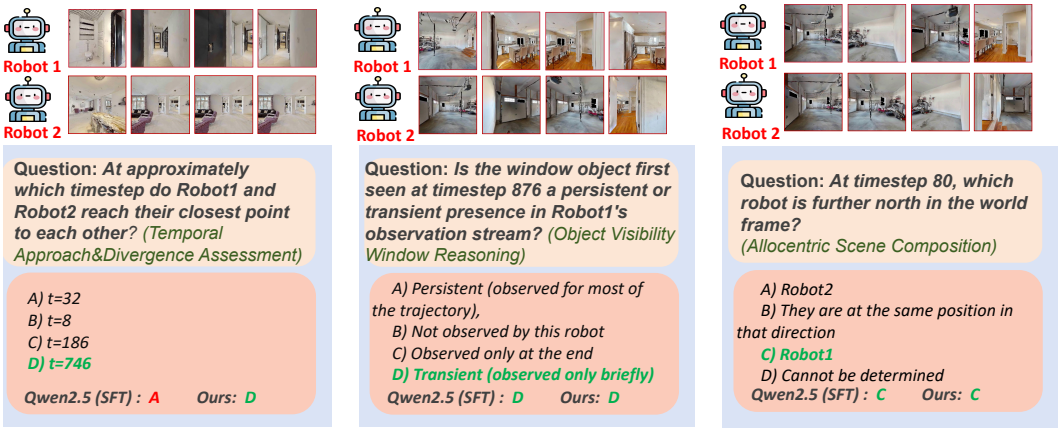


Figure 8: Overview of qualitative results of our approach and SFT Qwen2.5-VL-7B [24].

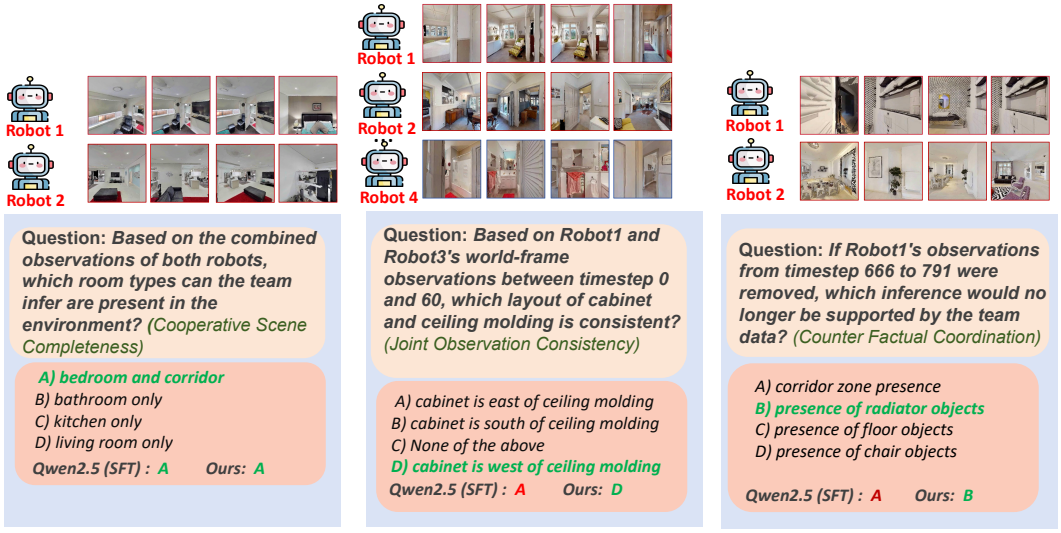


Figure 9: Overview of qualitative results of our approach and SFT Qwen2.5-VL-7B [24].

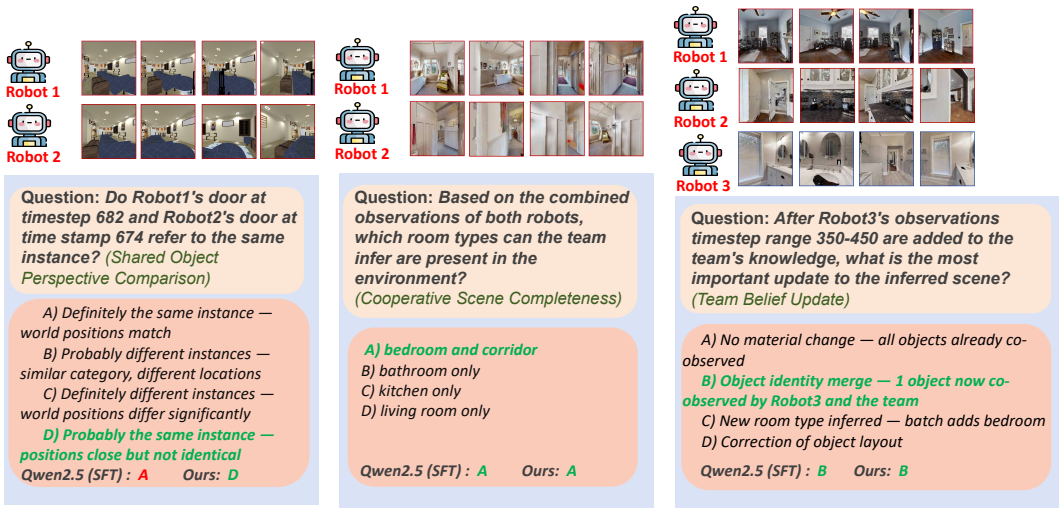


Figure 10: Overview of qualitative results of our approach and SFT Qwen2.5-VL-7B [24].

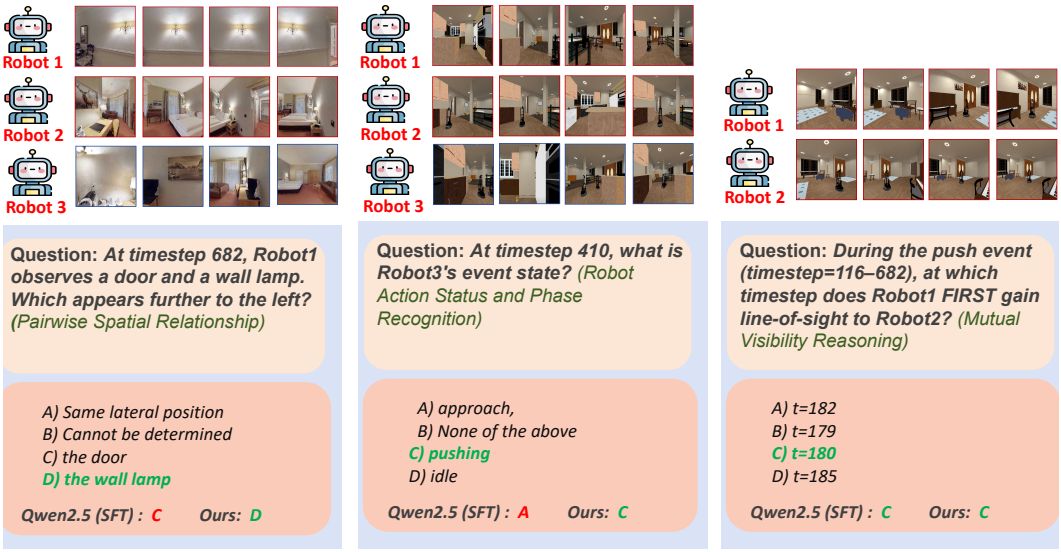


Figure 11: Overview of qualitative results of our approach and SFT Qwen2.5-VL-7B [24].

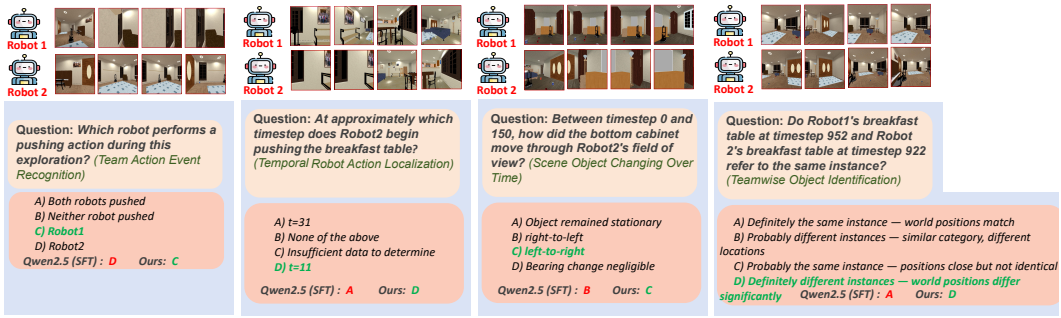


Figure 12: Overview of qualitative results of our approach and SFT Qwen2.5-VL-7B [24].

and team sizes. These results suggest that exchanging structured textual evidence can support cooperative reasoning more effectively than aggregating independent answer predictions. Nevertheless,

Table 9: Comparison of centralized and decentralized approaches on Habitat, measured by accuracy (%). All methods use Qwen2.5-VL-7B. T1–T4 denote reasoning tiers, and N denotes team size. Decentralized methods combine locally generated reports or predictions rather than jointly processing all videos. Best results are in bold.

		Reasoning tier				Team size		
Setting	Method	T1	T2	T3	T4	$N = 2$	$N = 3$	$N = 4$
Centralized	Zero-shot	23.57	39.61	28.40	34.35	34.66	34.99	31.03
Decentralized	Confidence-Weighted Vote	24.44	38.98	26.79	27.89	31.79	33.94	25.78
Decentralized	Structured Report Aggregator	31.11	48.02	30.36	47.81	38.97	44.34	42.67
Centralized	SFT	46.67	76.30	53.02	72.53	75.04	63.82	61.77
Centralized	EgoRAG	48.31	74.93	51.80	71.36	73.51	63.61	60.37
Centralized	TSPO	37.97	75.80	46.68	68.55	69.28	61.44	57.94
Centralized	**SG-CoR**	**55.36**	**78.72**	**59.36**	**74.94**	**78.26**	**67.34**	**66.60**

SG-CoR exceeds Structured Report Aggregator by 27.13%–30.70% on T2–T4 and 23.00%–39.29% across team sizes.

One possible limitation of report-based aggregation is that converting each stream into text before cross-robot comparison may discard fine-grained visual, temporal, or geometric evidence. Such information is relevant to relative pose, mutual visibility, cross-view object identity, observation consistency, and team-belief reasoning. Language communication therefore offers a complementary approach to cooperative reasoning, while the results motivate further investigation of representations that preserve cross-view evidence. These comparisons evaluate the reported systems; isolating the effect of communication format would require matched training and inference budgets.

Preliminary failure-case analysis. Qualitative inspection of **SG-CoR** errors identifies three recurring limitations. *First, insufficient visual evidence:* severe occlusion or small, distant objects can make the relevant evidence difficult to recover across all robot views. This motivates occlusion-aware evidence retrieval and uncertainty-based abstention when the available observations remain insufficient. *Second, incomplete temporal coverage:* the fixed frame budget can miss brief events or precise action transitions when decisive evidence occurs between selected frames, motivating adaptive temporal sampling. *Third, cross-view association errors:* visually similar robot streams and contradictory observations can still lead to agent-identity confusion or incorrect object associations, motivating stronger identity-preserving cross-view alignment. These observations provide qualitative directions for improvement; they do not constitute a quantitative breakdown of failure frequencies.

A.9 Contributions of Visual, Role, and Privileged Pose Cues

SG-CoR performs inference using robot-indexed egocentric RGB videos and the textual query, with the active robot count provided as structural metadata. Visual features provide appearance and spatial evidence, spectral features summarize temporal changes, and learned role embeddings encode episode-local robot identities. Motion and pose information provide privileged supervision during training through GMIPSD and are not required at inference.

To examine these contributions, we evaluate variants that remove the role token, visual token, or training-time pose token. Table 10 reports results using the Qwen2.5-VL-7B backbone on Habitat and iGibson. The pose-token ablation concerns privileged training information, rather than a test-time pose input.

Visual evidence. Removing the visual token produces the largest decrease among the evaluated variants: overall accuracy falls by 27.27% on Habitat and 28.54% on iGibson. This highlights the importance of direct visual evidence for cooperative spatial reasoning. Together with the spectral-energy and Fourier-embedding ablations in Table 3, these results support the complementary contributions of visual and spectral features.

Robot identity. Removing the role token reduces overall accuracy by 5.80% on Habitat and 7.94% on iGibson. These decreases support the value of explicit robot-identity cues during multi-robot fusion, beyond organizing the input streams by robot index.

Table 10: Ablation of role, visual, and training-time pose tokens using Qwen2.5-VL-7B. Accuracy (%) is reported across team sizes on Habitat and iGibson. AVG denotes accuracy over all evaluated QA samples. Best results are in bold.

	Habitat				iGibson			
Method	$N=2$	$N=3$	$N=4$	AVG	$N=2$	$N=3$	$N=4$	AVG
w/o role token	72.55	62.98	59.32	64.75	60.38	65.14	63.16	62.88
w/o visual token	33.37	47.42	48.24	43.28	35.68	45.36	45.70	42.28
w/o pose token (training)	72.15	63.06	59.42	64.69	59.93	64.85	62.47	62.40
SG-CoR	**78.26**	**67.34**	**66.60**	**70.55**	**72.74**	**69.92**	**69.85**	**70.82**

Table 11: Evaluation on seen scenarios using the Qwen2.5-VL-Instruct backbone. Three of the fifteen explorations per scenario are reserved for testing. Accuracy (%) is reported across robot team sizes, and AVG denotes accuracy over all evaluated QA samples. Best results are in bold.

	Habitat				iGibson			
Method	$N=2$	$N=3$	$N=4$	AVG	$N=2$	$N=3$	$N=4$	AVG
SFT (Qwen2.5-VL-Instruct)	61.83	61.03	55.69	59.43	53.42	67.61	53.95	58.18
SG-CoR (Qwen2.5-VL-Instruct)	**76.08**	**66.47**	**63.46**	**68.46**	**70.07**	**73.38**	**62.34**	**68.41**

Privileged training information. Removing the training-time pose token reduces overall accuracy by 5.86% on Habitat and 8.42% on iGibson. These results support the benefit of privileged pose supervision for inference without pose measurements. Together with the GMIPSD ablation in Table 3, they indicate that training-time geometric information can improve reasoning from egocentric visual observations.

A.10 Evaluation on Seen Scenarios

To complement the scene-disjoint evaluation, we additionally evaluate **SG-CoR** on held-out explorations within seen scenarios. For each scenario, three of the fifteen exploration episodes are reserved for testing. The split therefore operates at the exploration level: training and testing share scenarios, while the test set contains held-out exploration episodes. This setting evaluates reasoning over new observations within environments encountered during training.

As shown in Table 11, **SG-CoR** outperforms SFT across both datasets and all evaluated team sizes. Its overall accuracy reaches 68.46% on Habitat and 68.41% on iGibson, exceeding SFT by 9.03% and 10.23%, respectively. Across individual team sizes, the improvements range from 5.44% to 14.25% on Habitat and from 5.77% to 16.65% on iGibson. These results show that **SG-CoR**'s performance advantage extends to held-out explorations within seen scenarios, complementing the evaluation on unseen scenes.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The authors confirmed that the paper's contributions and scope are reflected in the abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitation is discussed in A.1.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Implementation details are provided and the methodology is described. The code will be publicly released upon decision.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The dataset and code are provided in the link at the end of the abstract.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The training details are shown in the implementation details section. The dataset splits are shown in Section 3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We verified our approach on both of the dataset collected from simulator and the dataset collected from the real-world lab.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details are mentioned in the implementation details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have carefully checked the code of ethics of NeurIPS and follow the guideline.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Society impact discussion is provided in the appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: We do not have this risk.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the assets are carefully cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide detailed description of our contributed dataset in both of the main paper and the appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: No crowdsourcing experiments and research with human subjects are conducted.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not constitute human subjects. The benchmark is developed from existing publicly available datasets, and the human verification component is only limited to internal annotation quality checks.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We use GPT5-mini to conduct the closed-source MLLM evaluation on the real-world evaluation set.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.