Unleashing Spatial Reasoning in Multimodal Large Language Models via Textual Representation Guided Reasoning

Jiacheng Hua[1,2] **Yishu Yin**[1] **Yuhang Wu**[1]
Tai Wang[2] **Yifei Huang**[3,2] **Miao Liu**[1†]
[1]College of AI, Tsinghua University, Beijing, China
[2]Shanghai Artificial Intelligence Laboratory, Shanghai, China
[3]The University of Tokyo, Tokyo, Japan
hjc21@mails.tsinghua.edu.cn miaoliu@mail.tsinghua.edu.cn
https://trace-reasoning.github.io

Abstract

Existing Multimodal Large Language Models (MLLMs) struggle with 3D spatial reasoning, as they fail to construct structured abstractions of the 3D environment depicted in video inputs. To bridge this gap, drawing inspiration from cognitive theories of allocentric spatial reasoning, we investigate how to enable MLLMs to model and reason over text-based spatial representations of video. Specifically, we introduce *Textual Representation of Allocentric Context from Egocentric Video (TRACE)*, a prompting method that induces MLLMs to generate text-based representations of 3D environments as intermediate reasoning traces for more accurate spatial question answering. TRACE encodes meta-context, camera trajectories, and detailed object entities to support structured spatial reasoning over egocentric videos. Extensive experiments on VSI-Bench and OST-Bench demonstrate that TRACE yields notable and consistent improvements over prior prompting strategies across a diverse range of MLLM backbones, spanning different parameter scales and training schemas. We further present ablation studies to validate our design choices, along with detailed analyses that probe the bottlenecks of 3D spatial reasoning in MLLMs.

1 Introduction

Cognitive science studies suggest that human reasoning about the 3D world relies on cortical mechanisms that transform visual input into hierarchical representations of objects and spatial relations, rather than operating directly on pixel-level stimuli (Marr and Nishihara, 1978). For instance, when humans approach the spatial reasoning question shown in Fig. 1(a), the solving process does not simply involve searching for cues within individual egocentric frames. Instead, we construct an immersive allocentric representation of the scene (Klatzky, 1998), mentally situating ourselves within the environment and reasoning about the underlying room layout to complement egocentric observations. Moreover, such allocentric representations can be vividly described using text alone, as demonstrated in Fig. 1(b). This observation naturally motivates the design of effective text-based video representations to enhance the spatial reasoning capabilities of existing Multimodal Large Language Models (MLLMs).

Recent studies show that existing MLLMs struggle with 3D spatial question answering (QA) (Yang et al., 2025b; Lin et al., 2025; Yang et al., 2025a), despite being pretrained on massive video datasets that inherently encode rich spatial information. One key reason is that these models often overly rely on 2D visual signals and learn spurious shortcut correlations from implicit spatial cues, rather than building hierarchical abstractions of the 3D scene. In this context, we raise a fundamental scientific question: Can MLLMs be guided to explicitly construct and reason over structured allocentric representations of 3D spatial environments from 2D visual observations?

Previous work on spatially aware MLLMs generally falls into two main directions: 1) curating large-scale supervised fine-tuning data for spatial reasoning QA (Daxberger et al., 2025; Ray et al., 2024), which limits scalability and generalization; or 2) incorporating additional geometric or stereo modalities into MLLMs (Cheng et al., 2024; Zhu et al., 2024), which increases system complexity and restricts applicability to off-the-shelf MLLMs. Our work explores a distinct formulation: inspired by prior approaches that extract textual descrip-

† Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13616–13637
July 2-7, 2026

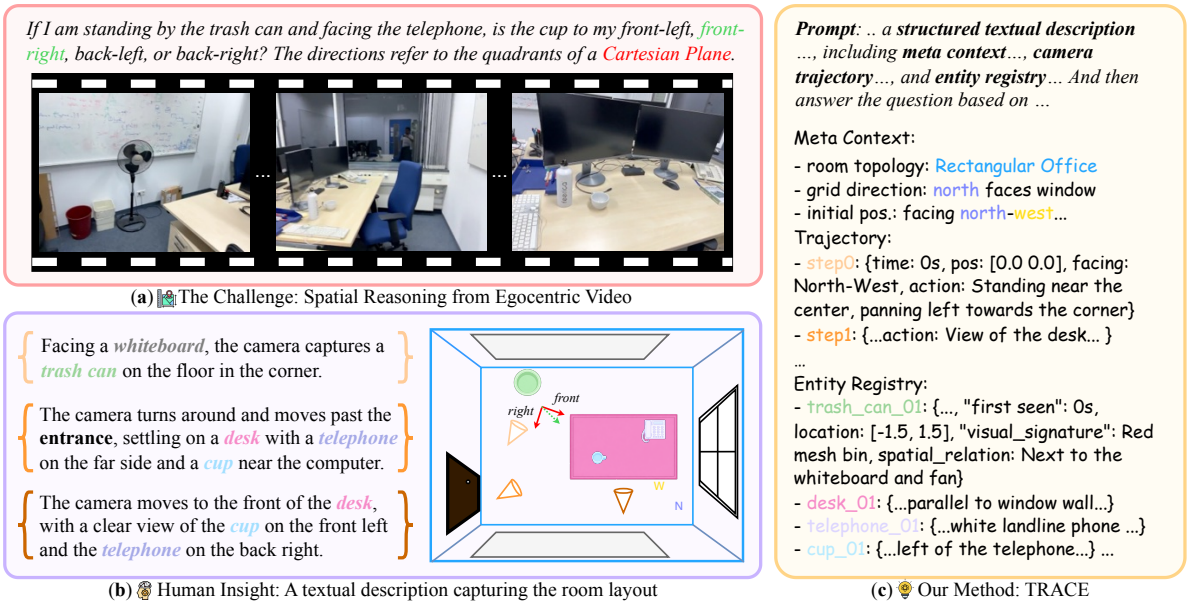

Figure 1: *Motivation for Textual Representation of Allocentric Context from Egocentric Video (TRACE) in video-based spatial reasoning.* (a) An egocentric video paired with a query that requires holistic spatial reasoning. (b) A textual description that vividly captures the room layout needed to solve the example spatial question answering (QA). (c) TRACE encodes meta-context, camera trajectory, and entities, serving as an intermediate reasoning trace for spatial QA with MLLMs.

tions from images or videos and then leverage only LLMs for VQA (Wang et al., 2024c; Fan et al., 2025), as well as Chain-of-Thought prompting methods (Wei et al., 2022), we propose to employ textual descriptions of 3D spatial structure as an intermediate reasoning trace that enables structured spatial reasoning in MLLMs.

Specifically, we introduce **TRACE**, short for **T**extual **R**epresentation of **A**llocentric **C**ontext from **E**gocentric Video, a prompting method that encourages MLLMs to generate a text-based allocentric representation of the 3D environment, facilitating spatial reasoning over the input egocentric video. As illustrated in Fig. 1(c), our proposed TRACE adopts a structured design that integrates *Meta Context* describing the room layout and coordinate system, camera *Trajectory* sampled over temporal windows, and explicit object *Entity Registry*. This design encourages MLLMs to perform explicit reasoning over a structured allocentric representation of the scene prior to answer generation.

We conduct extensive experiments on VSI-Bench (Yang et al., 2025a) and OST-Bench (Lin et al., 2025) to evaluate TRACE, demonstrating clear performance gains over prior prompting strategies. Comparisons with other text-based video spatial representations further validate the effectiveness of our approach. We also perform detailed ablation studies and decompositional analyses to probe the bottlenecks of 3D spatial reasoning. These results highlight structured textual allocentric representations as an effective intermediate reasoning interface for video-based spatial QA in MLLMs.

2 Related Work

Spatial Representation Prior work has extensively studied spatial reasoning with vision–language models (Johnson et al., 2017; Yang et al., 2019; Hudson and Manning, 2019). In addition, a significant body of work has examined vision–language models in embodied or navigation-oriented settings (Anderson et al., 2018; Chen et al., 2019; Shridhar et al., 2020). More recent work seeks to augment vision–language models with explicit 3D or geometric modalities (Hong et al., 2023; Cheng et al., 2024; Zhu et al., 2024), or with instruction tuning using carefully constructed data pipelines (Chen et al., 2024; Daxberger et al., 2025; Ray et al., 2024). Meanwhile, several diagnostic studies highlight that, despite these advances, current MLLMs still struggle to internally organize spatial information, motivating representations that more explicitly expose scene structure to the model (Wang et al., 2024a; Liao et al., 2024).

Our work is most closely related to recent efforts that investigate 3D spatial reasoning in MLLMs

through the lens of *intermediate representations for capturing scene structure* (Yang et al., 2025a; Wang et al., 2024a). Thinking in Space (Yang et al., 2025a) shows that explicitly externalizing a spatial representation—such as a cognitive map—can substantially improve spatial reasoning performance, whereas standard chain-of-thought prompting alone provides limited benefit. Complementarily, SpatialEval (Wang et al., 2024a) reveals that even strong multimodal LLMs often fail to construct consistent internal 3D representations and instead rely on shortcut correlations inherited from 2D pretraining. In contrast to introducing new geometric inputs, architectural modules, or large-scale spatial instruction tuning, we propose a text-based spatial representation that serves as an intermediate reasoning step to enhance the spatial reasoning capabilities of MLLMs. Hence, our approach is flexible and broadly applicable to off-the-shelf MLLMs.

Text-based Description of Video Textual description generation for video sequences has been extensively studied. Early models addressed video captioning using sequence-to-sequence and CNN-RNN architectures (Venugopalan et al., 2015; Donahue et al., 2015); later efforts focused on dense event captioning and paragraph-level video storytelling (Krishna et al., 2017; Li et al., 2018; Wang et al., 2021); another direction explored large-scale video-language pretraining for downstream tasks like retrieval and QA (Sun et al., 2019; Luo et al., 2020; Xu et al., 2021; Lei et al., 2021; Zhao et al., 2023; Yang et al., 2023).

Our work is more closely related to approaches that build structured textual representations of video content for LLM–based question answering (Wang et al., 2024c,b; Huang et al., 2025; Ren et al., 2025; Li et al., 2025; Kahatapitiya et al., 2025). These approaches treat linguistic descriptions as the primary medium for long-context video comprehension, rather than reasoning directly over raw frames. VideoTree (Wang et al., 2024c) builds a query-adaptive hierarchical tree of video segments and associated captions to support long-video QA with LLMs. VideoAgent (Wang et al., 2024b) uses an LLM as an agent to iteratively select informative clips/frames and maintain a running textual state for long-form video understanding. Video Mind Palace (Huang et al., 2025) constructs environment-grounded semantic graphs from videos as a persistent memory structure that an LLM can read for long-range reasoning. Instead of optimizing evidence coverage and retrieval over long temporal contexts, we focus on designing textual representations that enable MLLMs to explicitly reason over 3D geometry cues.

Prompting in M/LLM Prompting has become a primary inference-time mechanism for steering large M/LLM, including (i) rationale-based reasoning prompts (Wei et al., 2022; Kojima et al., 2022) (ii) decomposition and planning prompts that solve problems via sub-goals (Khot et al., 2022; Zhou et al., 2023; Wang et al., 2023a; Press et al., 2023) (iii) aggregation and search-style prompts to reduce variance and explore alternatives (Wang et al., 2023b; Yao et al., 2023; Besta et al., 2024) (iv) iterative self-improvement prompts via reflection (Gou et al., 2023). Another view is to treat language as an interface to external resources, using tool-augmented prompts and retrieval-mediated learning (Yao et al., 2022; Schick et al., 2023; Press et al., 2023; Trivedi et al., 2023). Inspired by prior work, we propose *TRACE*, the first prompting-based method that unleashes the spatial reasoning capability of MLLMs.

3 Method

Standard prompting methods, such as Chain-of-Thought (CoT) (Wei et al., 2022), encourage Multimodal Large Language Models (MLLMs) to generate intermediate reasoning steps to bridge the gap between input and output. While effective for arithmetic and symbolic tasks (Cobbe et al., 2021; Hudson and Manning, 2019), standard chain of thought and other linguistic prompting strategies often fall short or even hurt performance on complex spatial reasoning tasks (Yang et al., 2025a). Our key intuition is that MLLMs may need to explicitly reason over an intermediate global representation of the 3D scene to complement the egocentric video inputs used in most spatial intelligence benchmarks.

To this end, drawing inspiration from human cognitive processes (Marr and Nishihara, 1978), we introduce **Textual Representation of Allocentric Context from Egocentric Video** (**TRACE**), a method that encourages MLLMs to generate a text-based allocentric representation of the 3D environment that facilitates spatial question answering. In the following sections, we first introduce the problem setting of spatial question answering with prompting, then describe the key components of our TRACE design, and finally elaborate on the

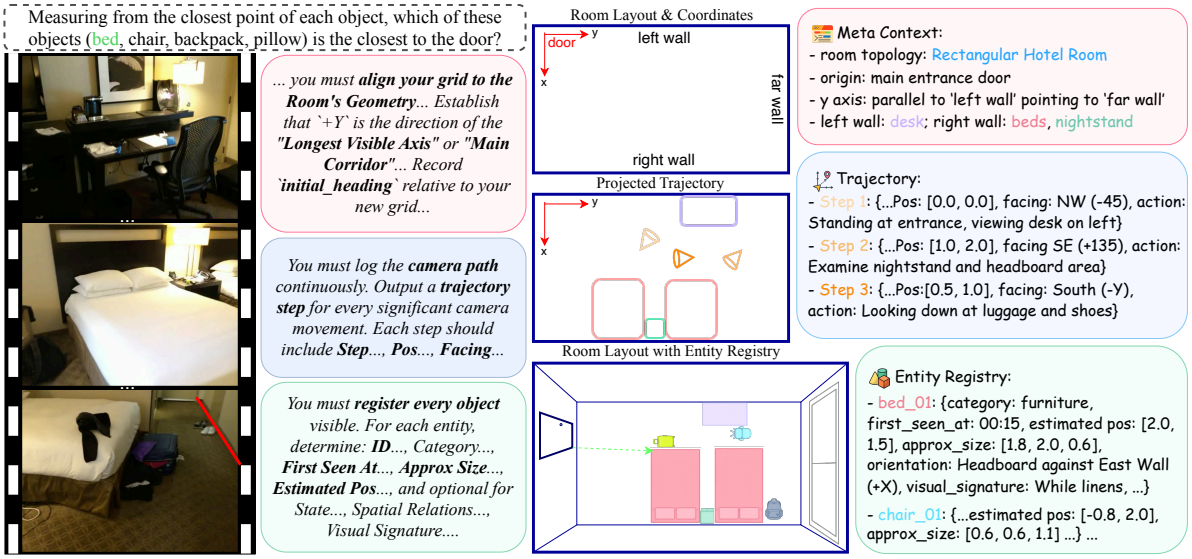

Figure 2: *Illustration of our Textual Representation of Allocentric Context from Egocentric Video (TRACE).* We construct TRACE by aligning a global coordinate system with the room layout and geometry, logging the camera trajectory across temporal steps, and registering visible objects with key attributes, estimated positions, and spatial relations. Here, we also show the key prompts used to guide MLLMs to generate this intermediate reasoning trace.

inference schema.

3.1 Problem Formulation

We formulate spatial reasoning as a generation task conditioned on a given egocentric video $V = \{v_1, ..., v_T\}$ and a natural language query Q, with the objective of generating the answer A.

Standard CoT approaches model the probability $\mathrm{P}(A, R|V, Q)$, where R is a specified reasoning trace. However, previous reasoning traces often fail to capture the geometric structure required for spatial tasks (Yang et al., 2025b). We instead enforce a protocol where the reasoning trace takes the form of a Textual Representation of Allocentric Context from Egocentric Video, denoted as $\mathcal{G}$. The inference process is formalized as a single-turn generation maximizing:

$$\hat{A}, \hat{\mathcal{G}} = \underset{A,\mathcal{G}}{\operatorname{argmax}} \underbrace{P(A|\mathcal{G}, V, Q)}_{\text{Reasoning Parser}} \cdot \underbrace{P(\mathcal{G}|V, Q)}_{\text{Spatial Descriptor}}$$

Here, the Spatial Descriptor produces intermediate reasoning steps as TRACE, which the Reasoning Parser then uses to generate the final answer.

3.2 Key Components of TRACE

We formally define TRACE as a tuple $\mathcal{G} = \langle \mathcal{M}, \mathcal{T}, \mathcal{E} \rangle$. Here, $\mathcal{M}$ denotes the meta context, including the room topology, grid alignment, and the observer's initial heading. The trajectory $\mathcal{T} = \{(t_k, p_k, \phi_k)\}_{k=0}^{K}$ records the observer's position $p_k \in \mathbb{R}^d$ and heading ϕ_k at discrete time steps t_k. Finally, $\mathcal{E} = \{e_j\}_{j=1}^{N}$ denotes the set of N entities.

Meta Context A common failure mode in spatial reasoning arises from losing track of camera initialization and the corresponding coordinate system. We propose a Room Aligned Coordinate System that is initialized from a coarse room layout sketch, for example a rectangular bedroom. We fix the origin $[0, 0]$ at starting position of the observer, and then establishes the y axis by detecting the most salient straight line characterized by static large objects rather than the camera's initial heading.

Camera Trajectory Static maps fail to capture the dynamic nature of video. To address this limitation, we require the model to reconstruct the observer path as a discrete sequence of steps using the established coordinate system and large static objects from the Meta Context $\mathcal{M}$ as reference points. For each step, TRACE records the timestamp, estimated position $[x, y]$, and the camera's facing direction. We approximate camera direction using eight discrete orientations defined by the cardinal directions, with the y axis aligned with north, as accurate numerical angle estimation is difficult for the Scene Descriptor and continuous pose representations pose challenges for the Reasoning Parser. In addition, we include an action property that encodes the camera centric motion context. Our formulation thus effectively reconstructs the surveyor's path, allowing the model to

answer navigation and route-planning questions by traversing the generated static map rather than relying on transient visual memory.

Entity Registry Instead of predicting loose grid cells as in the Cognitive Map (Yang et al., 2025a), our model maintains a registry of observed entities with detailed attributes throughout the temporal sequence. To prevent object duplication and ensure precise localization, we enforce a structured schema for each object entity:

- *Temporal Stamping:* Each entity e_i must include a timestamp recording its first seen time, aiding in object tracking.
- *Visual Signature:* Each entity includes a brief appearance based description that captures its salient visual attributes, which helps disambiguate visually similar instances across time.
- *Metric Estimation:* TRACE records plausible 2D coordinates $[x, y]$ in meters for every entity relative to the grid origin. While these coordinates are estimates, the act of estimation forces the model to resolve spatial relations (e.g., *near*, *between*) into geometric constraints.
- *Spatial Relations:* Each entity records its relative spatial relations to nearby entities using natural language, providing complementary relational cues beyond absolute coordinates.
- *Strict Serialization:* Entities should be listed individually (e.g., `chair_01`, `chair_02`) rather than grouped, ensuring granular counting and positional accuracy.

3.3 Inference Mechanism

The inference of our standard implementation is performed in a single pass. We condition the generation process to explicitly yield the schema-compliant representation $\mathcal{G}$ prior to the final response. This acts as a structured Chain-of-Thought, where the generation $\mathcal{G}$ effectively loads the context window with a "spatial cache" of the environment. The final answer is then derived via TRACE-conditioned inference, which jointly accounts for the egocentric video input and queries the cached TRACE to compute Euclidean distances between objects coordinates $\mathcal{E}$ or traverse nodes in $\mathcal{T}$. This mechanism improves final answer accuracy by grounding answer generation in previously generated and verifiable geometric constraints.

4 Experiments

4.1 Experimental Setup

Benchmarks We consider two spatial intelligence related benchmarks: VSI-Bench (Yang et al., 2025a) and OST-Bench (Lin et al., 2025).

VSI-Bench is a video-based benchmark built from egocentric indoor scene scans, containing 5,130 question-answer (QA) pairs across 288 real-world videos. It covers eight tasks spanning configurational, measurement-estimation, and spatiotemporal reasoning. In contrast, OST-Bench assesses online spatio-temporal understanding from the perspective of an embodied agent actively exploring a scene. Comprising 1,386 scenes and 10,165 QA pairs, it employs a multi-round dialogue format that requires models to process incrementally acquired observations and integrate historical memory to answer questions regarding the agent's state, visible information, and spatial relationships. In this work, we evaluate on the full set of VSI-Bench, while for OST-Bench, we use a reproducible random subset consisting of 200 scenes and 1,396 QA pairs.

Metrics Current Spatial AI benchmarks mainly follow two formats: multiple-choice questions (MCQ) and numerical questions. For MCQ, we report Accuracy (Acc). To evaluate model predictions, we extract the answer option using exact matching, supplemented by fuzzy matching to robustly handle variations in model output formats (e.g., capturing the option letter or full text).

For numerical questions, we adopt the Mean Relative Accuracy (MRA) introduced by Yang et al. (2025a). MRA quantifies the proximity of a predicted value $\hat{y}$ to the ground truth y by averaging performance across a range of strictness thresholds $\mathcal{C} = \{0.5, 0.55, \ldots, 0.95\}$. MRA is formally defined as:

$$\text{MRA} = \frac{1}{|\mathcal{C}|} \sum_{\theta \in \mathcal{C}} \mathbb{I}\left(\frac{|\hat{y} - y|}{y} < 1 - \theta \right)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. A prediction is considered correct at threshold θ only if its relative error is less than $1 - \theta$.

Model Selection We validate the effectiveness of our approach using Gemini 3 Pro (Gemini Team, 2025) as our primary proprietary model. All open-source baselines are evaluated using their default configurations and parameters. For VSI-Bench, we report main results using both Qwen2.5-VL-72B (Bai et al., 2025) and MiMo-VL-7B (Xiaomi,

Table 1: *Evaluation results on the VSI benchmark.* We report average performance and detailed breakdowns across numerical-answer and multiple-choice tasks, under proprietary and open-sourced base models. Best results are in **bold**, and second-best are underlined.

Methods	Avg.	Numerical Answer				Multiple-Choice Answer			
		Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route	Order
Gemini 3 Pro as base model									
Direct	52.61	33.77	32.57	67.09	42.99	62.54	50.52	51.03	70.71
CoT	53.65	30.35	34.54	64.05	40.76	61.78	58.09	**61.34**	71.96
ToT	58.88	44.55	**42.12**	72.20	45.55	65.35	57.83	55.62	**73.73**
LtM	59.52	45.19	40.72	73.36	44.15	**65.82**	60.40	53.59	73.64
CM	59.72	46.70	41.43	72.49	**50.14**	63.69	58.62	55.50	72.61
Ours	**60.15**	**47.55**	38.82	**73.90**	45.62	63.85	**61.70**	58.01	72.97
Qwen2.5-VL-72B-Instruct as base model									
Direct	36.28	**33.36**	20.53	49.31	41.49	**43.38**	27.79	32.47	**44.01**
CoT	29.78	21.27	24.95	16.31	40.94	39.44	33.16	28.87	43.53
ToT	38.06	17.89	26.20	53.15	**47.01**	41.55	36.78	**35.05**	**44.01**
LtM	38.01	23.27	**31.39**	54.49	38.68	42.96	34.71	29.90	36.73
CM	35.47	21.58	15.67	52.65	37.26	39.44	36.05	34.54	42.39
Ours	**39.38**	22.05	28.03	**59.98**	38.99	40.85	**37.40**	31.96	42.56
MiMo-VL-7B as base model									
Direct	39.79	**36.02**	29.84	52.38	42.95	40.14	33.78	31.44	47.41
CoT	37.49	34.27	23.50	48.52	43.23	38.73	32.75	27.84	49.23
ToT	39.14	29.45	30.44	54.26	40.14	41.41	32.02	32.47	46.60
LtM	38.34	35.09	24.47	48.22	**44.48**	**43.10**	30.79	**35.05**	49.50
CM	36.85	27.43	23.14	50.14	39.06	41.41	32.54	27.84	46.76
Ours	**41.42**	33.27	**31.51**	**58.67**	41.56	39.44	**35.33**	28.87	**51.29**

Table 2: *Evaluation results on the OST benchmark.* Results are reported across agent state understanding, visible information reasoning, and agent–object spatial relationship tasks, under proprietary and open-sourced base models. Best results are in **bold**, and second-best are underlined.

Methods	Avg.	Agent Visible Info					Agent-object Spatial Relationship						Agent State			
		Exist.		Quant.	Divers.	Order	Direct.			Dist.			Pos.		Orient.	
		JUD.	TEMP.	CNT.	JUD.	JUD.	JUD.	TEMP.	EST.	JUD.	TEMP.	EST.	JUD.	EST.	JUD.	EST.
Gemini 3 Pro as base model																
Direct	59.22	**96.72**	84.87	**68.75**	89.66	**82.54**	54.27	48.15	28.63	**60.61**	54.55	30.00	71.43	22.78	72.60	22.68
CoT	59.26	82.24	**93.99**	65.96	96.55	77.78	52.24	**62.96**	27.84	52.76	50.00	31.64	71.43	24.26	72.60	26.67
ToT	59.20	94.54	83.55	66.67	93.10	80.65	54.27	54.39	31.00	55.61	53.85	**32.36**	61.76	29.07	72.60	24.58
LtM	59.27	95.65	85.52	66.67	**100.0**	76.19	53.00	50.91	25.88	55.78	53.03	31.23	71.43	**35.85**	69.86	18.06
CM	59.04	95.05	86.18	**68.75**	89.66	77.78	55.72	48.15	25.40	60.30	49.23	26.30	**80.00**	24.81	69.86	28.47
Ours	**60.42**	95.05	86.58	66.67	96.43	77.78	**57.00**	52.73	**34.12**	60.30	**56.45**	31.25	54.29	31.92	**76.71**	**29.03**
MiMo-VL-7B as base model																
Direct	62.65	**92.39**	51.63	53.06	**100.0**	85.71	63.68	21.05	34.12	76.38	40.91	**28.77**	**100.0**	9.26	91.78	40.28
CoT	61.69	89.67	48.37	48.98	**100.0**	82.54	68.66	15.79	33.33	75.38	39.39	**28.77**	97.14	11.30	89.04	39.31
ToT	62.20	91.30	52.29	51.02	**100.0**	**90.48**	64.68	15.79	29.22	75.38	40.91	27.53	**100.0**	15.74	84.93	**41.39**
LtM	63.75	88.04	44.44	**63.27**	**100.0**	88.89	**77.11**	21.05	**35.88**	76.88	40.91	22.33	**100.0**	7.78	**100.0**	36.94
CM	64.00	88.59	54.90	57.14	**100.0**	88.89	69.15	21.05	33.92	**78.89**	36.36	27.53	**100.0**	11.85	97.26	38.89
Ours	**65.04**	91.85	**57.52**	61.22	**100.0**	87.30	69.15	**24.56**	32.35	74.87	**42.42**	26.44	**100.0**	**29.07**	94.52	38.06

2025). Additional experiments on VSI-Bench with other state-of-the-art models, including o3 (OpenAI, 2025) and GLM-4.5V (V Team et al., 2025), are detailed in Appendix B. For OST-Bench, we adopt MiMo-VL-7B as our open-source backbone, omitting the Qwen series due to its documented limitations in multi-turn instruction-following settings (Lee et al., 2025).

4.2 Experimental Results

Comparison of Different Prompting Methods We first contrast our method with previously proposed prompting methods that have demonstrated effectiveness on general VQA tasks. Specifically, we consider the following prompting strategies:

- *Chain-of-Thought (CoT)* (Wei et al., 2022): Elicits a step-by-step reasoning trace to bridge the gap between the input and the final answer.
- *Tree-of-Thought (ToT)* (Yao et al., 2023): Explores a tree of potential reasoning paths, evaluating and selecting the most promising intermediate thoughts to derive the answer.
- *Least-to-Most (LtM)* (Zhou et al., 2023): Decomposes complex spatial queries into manageable sub-problems, solving them sequentially to guide the final inference.
- *Cognitive Map (CM)* (Yang et al., 2025a): Instructs the model to construct a 10×10 semantic grid capturing the coarse layout of relevant objects before answering.

To ensure fair comparison and seamless integration of different prompting techniques, we keep

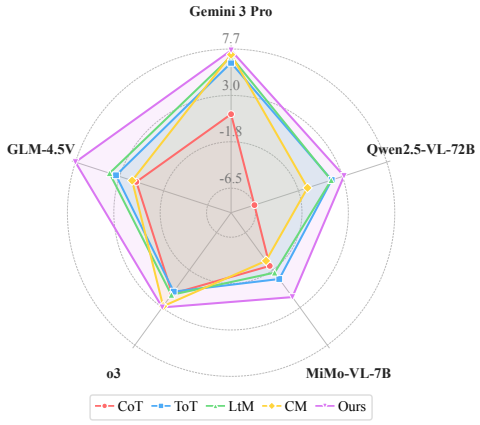

Figure 3: *Performance gains across models on VSI-Bench.* TRACE yields consistent, state-of-the-art performance gains compared to Direct prompting baselines, across various model architectures and parameter scales.

the prompting scaffold the same (e.g., identical input formatting, answer constraints, and post-processing), and vary only the method-specific instructions required by each prompting technique. We provide all prompts in Appendix A.

We evaluate our method, TRACE, alongside the Direct baseline and prior prompting strategies. Results are summarized in Tab. 1 and Tab. 2.

On VSI-Bench, advanced prompting methods consistently improve performance for Gemini, but yield only marginal gains or even compromise performance for Qwen. This discrepancy is likely due to the weaker instruction-following capability of the Qwen series, which limits its ability to effectively leverage prompting strategies for in-depth reasoning. Notably, our proposed TRACE yields substantial performance improvements of +7.54%, +3.10% and +1.63% for Gemini, Qwen and MiMo, respectively. These results demonstrate the robustness of our approach across different base models. In addition, we note that the latest Gemini 3 series incorporate step-by-step thinking instruction during training data construction, which likely leads to stronger alignment with existing prompting strategies and thus an inherent advantage. Even so, TRACE consistently outperforms these approaches on Gemini. Furthermore, additional experiments with other state-of-the-art models also demonstrate consistent performance gains with TRACE, as visualized in Fig 3.

On the OST benchmark, existing prompting strategies yield only marginal performance gains for both Gemini and MiMo models. This is because OST primarily evaluates multi-turn spatial reasoning, where step-by-step thinking prompts may hinder the model's ability to accurately ground and update spatial context across turns. In contrast, TRACE yields a +1.2% absolute performance gain on Gemini, and a +2.4% gain on the open-source MiMo. Notably, for the compact MiMo backbone, spatial specific prompting (CM and TRACE) prove superior to general linguistic reasoning (CoT, LtM and ToT), underscoring the effectiveness of explicit geometric grounding for smaller models. We do acknowledge, however, that TRACE can lead to a performance drop in certain agent state predictions. This limitation arises because TRACE is currently formulated as a static global allocentric representation. While this global perspective provides a highly stable environment model for relational reasoning, it creates a decoupling from the rapid, dynamic egocentric updates required for precise real-time agent state tracking.

Comparison of Different Prediction Setting We further examine how carefully designed text-based video representations can improve 3D spatial understanding. In particular, we explore the following prediction settings through which MLLMs can leverage such text-based representations:

- *One-Stage Inference* is the setup discussed in Sec. 3, where the model generates TRACE and answers the question using both the representation and the video input in a single pass.
- *Two-Stage Inference* first generates TRACE, which is then treated as additional context and fed into the MLLM, together with the video input, for final question answering.
- *Text-Only Inference* first generates our proposed TRACE and then uses an LLM to answer the question based solely on TRACE.

For fair comparison, we adopt the same MLLM and prompt components to construct TRACE representations across all settings. As shown in Tab. 3, the text-only approach achieves on-par performance with the direct video-based method using Gemini, suggesting that TRACE provides an informative summary of the video sequence. Another important finding is that, for both Qwen and Gemini, the one-stage prompting setting outperforms the two-stage setting. This suggests that not only is the resulting text-based representation beneficial, but the reasoning process involved in generating it also plays a critical role in enabling MLLMs to make accurate predictions.

Table 3: *Systematic studies of different prediction settings for utilizing our proposed text-based spatial representation.* Among all settings, one-stage prompting yields the best performance on both Gemini-3 Pro and Qwen2.5-VL-72B.

Input Setting	Avg.	Numerical Answer				Multiple-Choice Answer			
		Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route	Order
Proprietary model as base									
Video Direct	52.61	33.77	32.57	67.09	42.99	62.54	50.52	51.03	70.71
One-Stage	60.15	47.55	38.82	73.90	45.62	63.85	61.70	58.01	72.97
Two-Stage	58.52	42.25	36.73	72.10	52.17	58.75	63.50	51.35	74.01
Text-Only	52.27	28.52	32.66	67.28	48.02	49.58	62.66	52.43	64.93
Open-sourced model as base									
Video Direct	37.58	32.58	24.51	55.26	39.13	41.13	28.93	31.44	43.20
One-Stage	38.92	25.47	26.93	58.18	40.42	37.46	36.15	29.38	45.79
Two-Stage	32.85	16.80	19.75	42.33	26.46	37.32	34.19	34.02	45.95
Text-Only	31.11	12.83	21.74	39.71	23.51	37.04	33.16	32.99	40.13

Table 4: *Comparison with existing text-based spatial representations and ablation studies of our method.* We use Qwen2.5-VL-72B as base and adopt text-only inference for more direct comparison.

Input Setting	Avg.	Numerical Answer				Multiple-Choice Answer			
		Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route	Order
Cognitive Map	21.41	7.82	9.86	8.17	21.22	33.94	35.23	32.99	30.26
Spatial Caption	27.58	14.90	14.20	24.86	30.59	36.06	34.40	36.60	36.73
Ours	31.11	12.83	21.74	39.71	23.51	37.04	33.16	32.99	40.13
Ours w/o Trajectory	29.19	10.16	18.19	33.93	29.41	35.92	37.19	31.96	32.85
Ours w/o Entity Registry	25.87	6.11	28.84	7.41	19.69	41.69	37.50	30.41	33.50

Comparison with other Text-based Spatial Representations We compare our method with the most relevant cognitive map representation proposed in (Yang et al., 2025a) and a spatial captioning approach inspired by (Zhang et al., 2024), which sequentially describes the spatial components of the video sequence. To explicitly quantify the benefits of the text-based representation, we adopt the aforementioned text-only inference setting. As shown in Tab. 4, our method outperforms Cognitive Map by 9.7% and Spatial Caption by 3.53% on the VSI-Bench, highlighting the advantage of our proposed spatial representation. Furthermore, a visual comparison in Fig. 4 shows this advantage qualitatively, illustrating that TRACE captures the essential 3D granularity required for complex spatial reasoning that the cognitive map approach lacks.

Ablation Studies We ablate the key components of our method in Tab. 4. Removing trajectory information results in a 1.92% performance drop, while excluding entity registry leads to a larger drop of 5.24%, suggesting camera trajectory and entity registry play important roles in spatial QA. As expected, removing the entity registry leads to a substantial performance drop on object related tasks, while removing camera trajectory mainly affects performance on distance and order related reasoning. In addition, we find that removing trajectory information improves performance on room size and relative direction tasks. This suggests that current MLLMs lack the ability to reliably estimate camera motion, which can confuse models on tasks that require alignment-based reasoning.

Figure 4: *A visual illustration demonstrates that TRACE is more effective than the cognitive map (CM) approach.* Notably, the CM lacks the 3D granularity required for many spatial reasoning tasks.

4.3 Additional Analysis

Decomposing 3D Spatial Understanding Prior works (Yang et al., 2025b,a) have shown that existing MLLMs have limited 3D spatial understanding capabilities. We seek to provide an in-depth analysis of the underlying causes using a text only inference setting. Concretely, we decompose 3D

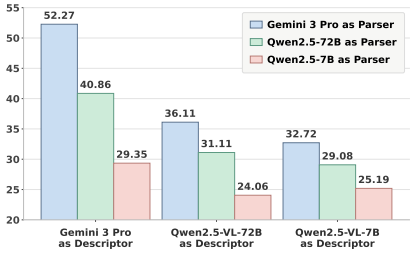

Figure 5: *Decompositional analysis of the reasoning parser and spatial descriptor.* The Qwen series lags behind the state-of-the-art Gemini 3 on both spatial reasoning and visual perception.

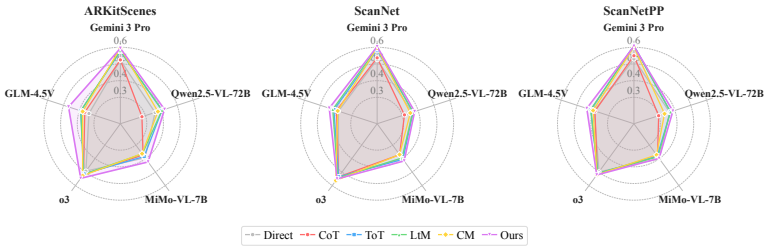

Figure 6: *Stratified analysis on VSI-Bench.* Our method (TRACE) consistently achieves robust performance gains across all granular scene distributions, demonstrating reliable generalization across diverse spatial layouts and complexities without aliasing to specific environment types.

reasoning into two stages: 3D visual perception and language-based spatial reasoning. Specifically, we use MLLMs as both a *Spatial Descriptor* for 3D grounding and a *Reasoning Parser* for spatial knowledge inference.

As shown in Fig. 5 using Gemini 3 Pro as a theoretical performance upper bound, we observe a significant performance drop when either the Descriptor or the Parser is replaced with Qwen2.5-VL-72B, especially when replacing the Descriptor. In addition, replacing the spatial descriptor from Qwen2.5-72B with Qwen2.5-7B results in only a marginal performance drop, whereas swapping the reasoning parser from Qwen2.5-72B to Qwen2.5-7B leads to a substantially larger degradation. This suggests that the two models exhibit comparable 3D visual perception capabilities, while the 72B variant has a markedly stronger reasoning capacity. Such decompositional analysis therefore helps identify the key bottlenecks in prevailing LLMs.

Token Efficiency We observe that the token length induced by thinking-based methods is highly sensitive to the choice of model backbone. Notably, our method achieves greater token efficiency while delivering better performance than advanced baselines (e.g., ToT and LtM) on compact models like MiMo, underscoring its strong potential for lower-latency embodied AI deployments. The same trend holds for specific large models, such as GLM, although our method is slightly more token-intensive on some other large foundation models. We refer readers to Appendix B for a more comprehensive breakdown and analysis. Importantly, optimizing token efficiency during the reasoning process constitutes a largely orthogonal research direction, which we leave for future work.

Cross-Environment Generalization To investigate whether TRACE is biased toward specific environment types, we stratify the VSI-Bench results across its core underlying datasets: ARKitScenes (Baruch et al., 2021), ScanNet (Dai et al., 2017), and ScanNetPP (Yeshwanth et al., 2023). These datasets represent a diverse range of indoor spatial layouts, scanning fidelities, and environmental complexities. As shown in Fig. 6, our structured prompting approach consistently delivers robust performance gains across all three scene distributions and five different model architectures. This confirms that our method's effectiveness is not restricted to a specific environment type, but rather generalizes reliably across varied spatial features and complexities.

5 Conclusions

We presented **TRACE**, a prompting approach that enables MLLMs to leverage the **T**extual **R**epresentation of **A**llocentric **C**ontext from **E**gocentric Video as an intermediate reasoning trace for spatial understanding. By explicitly modeling scene structure through meta-context, camera trajectory, and entity-level grounding, TRACE consistently improves performance on VSI-Bench and OST-Bench across diverse proprietary and open-source model backbones. Comparisons against prior linguistic prompting methods and other text-based spatial representations, together with detailed ablation studies, validate the effectiveness of our design choices. We further provide insights into how to effectively leverage text-based representations and present decompositional analyses that reveal common failure modes in MLLM spatial reasoning. More broadly, we hope TRACE can serve as a simple and widely applicable interface for studying structured spatial reasoning in off-the-shelf MLLMs. Overall, our results suggest a promising direction for advancing spatial reasoning in MLLMs and motivate further exploration of cognitively inspired representations.

Limitations and Future Work

Our work presents an initial attempt to design text-based representations that facilitate effective spatial reasoning in MLLMs. Nevertheless, our approach is still subject to several limitations. First, our current framework is formulated as a static allocentric representation. While this ensures global topological consistency, it somewhat creates a decoupling from the dynamic egocentric updates required for precise real-time agent state tracking in multi-turn settings. Furthermore, for the sake of fair comparison, our current implementation relies on the vision-language model itself to generate the representation. In practice, incorporating specialized visual expert models may further enhance the accuracy of the generated scene structures.

There are several promising directions for future work. A natural next step is to develop a dynamic streaming TRACE framework that incrementally updates the camera trajectory and entity registry as new observations arrive, allowing the model to maintain a persistent world model while recursively re-projecting the agent's pose within the map. Another important direction is to internalize TRACE-like reasoning into MLLMs through training, for example, by using TRACE to automatically construct high-quality spatial reasoning supervision for supervised finetuning, and potentially further improving the resulting policies with reinforcement learning. This could enable structured spatial representations to become part of the model's native reasoning process rather than an external prompt artifact. It would also be valuable to integrate specialized 3D perception modules into TRACE to improve the fidelity of the intermediate representation while preserving the flexibility of language-based reasoning. Future work could additionally study representation compression and token efficiency, since compact yet faithful spatial traces would be especially important for low-latency embodied agents. Finally, extending TRACE beyond QA to tasks such as navigation, planning, and manipulation may help establish whether structured textual world models can serve as a more general interface between perception and reasoning in multimodal systems.

Acknowledgments

This work was supported in part by Shanghai Artificial Intelligence Laboratory, the Zhiyuan Scholar Program of the Beijing Municipal Science and Technology Commission (Z251100008125045), NSFC Grants, and a research grant from the ByteDance Seed Team.

References

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

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

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and 1 others. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17682–17690.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12538–12547.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers

to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*.

Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, and 1 others. 2025. Mm-spatial: Exploring 3d spatial understanding in multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7395–7408.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. 2025. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pages 75–92. Springer.

Gemini Team. 2025. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 2023. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494.

Zeyi Huang, Yuyang Ji, Xiaofang Wang, Nikhil Mehta, Tong Xiao, Donghyun Lee, Sigmund Vanvalkenburgh, Shengxin Zha, Bolin Lai, Licheng Yu, and 1 others. 2025. Building a mind palace: Structuring environment-grounded semantic graphs for effective long video analysis with llms. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24169–24179.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Kumara Kahatapitiya, Kanchana Ranasinghe, Jongwoo Park, and Michael S Ryoo. 2025. Language repository for long video understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5627–5646.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406*.

Roberta L Klatzky. 1998. Allocentric and egocentric spatial representations: Definitions, distinctions, and interconnections. In *Spatial cognition: An interdisciplinary approach to representing and processing spatial knowledge*, pages 1–17. Springer.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.

Young-Jun Lee, Byung-Kwan Lee, Jianshu Zhang, Yechan Hwang, Byungsoo Ko, Han-Gyu Kim, Dongyu Yao, Xuankun Rong, Eojin Joo, Seung-Ho Han, and 1 others. 2025. Multiverse: A multiturn conversation benchmark for evaluating large vision and language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 708–719.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7331–7341.

Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. 2018. Jointly localizing and describing events for dense video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7492–7500.

Yiming Li, Xiaoshan Yang, Bing-Kun Bao, and Changsheng Xu. 2025. Graph prompts: Adapting video graph for video question answering. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 1485–1493.

Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, and David Acuna. 2024. Reasoning paths with reference objects elicit quantitative spatial reasoning in large vision-language models. *arXiv preprint arXiv:2409.09788*.

Jingli Lin, Chenming Zhu, Runsen Xu, Xiaohan Mao, Xihui Liu, Tai Wang, and Jiangmiao Pang. 2025. OST-bench: Evaluating the capabilities of MLLMs in online spatio-temporal scene understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*.

David Marr and Herbert Keith Nishihara. 1978. Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 200(1140):269–294.

OpenAI. 2025. Openai o3 and o4-mini system card.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, and 1 others. 2024. Sat: Dynamic spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*.

Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. 2025. Videorag: Retrieval-augmented generation with extreme long-context videos. *arXiv preprint arXiv:2502.01549*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 10014–10037.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 69 others. 2025. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Sharon Li, and Neel Joshi. 2024a. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems*, 37:75392–75421.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.

Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. 2021. End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6847–6857.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2024b. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2024c. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. *arXiv preprint arXiv:2405.19209*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

LLM-Core-Team Xiaomi. 2025. Mimo-vl technical report. *Preprint*, arXiv:2506.03569.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. 2023. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *CVPR*.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2025a. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643.

Kaiyu Yang, Olga Russakovsky, and Jia Deng. 2019. Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2051–2060.

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

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

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

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. 2024. A simple llm framework for long-range video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21715–21737.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. 2023. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. 2024. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*.

Appendix

This appendix provides technical details of our prompting method (Appendix A) and additional experiments and results (Appendix B).

A Prompting Details

We evaluate multiple prompting methods on our benchmark. Most methods share the same *base system prompt* and differ mainly in the *user prompt*. See more details below.

A.1 Overall Structure

Base system prompt. All methods (except the cognitive map method in §A.2) use the following base system prompt. This design allows us to control variables across prompting strategies: although the Cognitive Map and our TRACE method require specialized system instructions to define their intermediate representations, we keep these system prompts as close as possible to the base prompt (e.g., identical task framing and answer constraints) so that observed differences are primarily attributable to the prompting protocol rather than unrelated changes in instruction wording.

```
SYSTEM_PROMPT = """You are a multimodal
large language model being evaluated
on visual-spatial reasoning tasks with
egocentric indoor videos.

You are given:
- an egocentric video of an indoor
environment, and
- a question about that video.

Your goal is to answer the question as
accurately as possible.
```

```
Answer format:
- {POST_PROMPT}
- Do NOT add extra text on the final
answer line (no units, no explanations).
"""
```

Prompt assembly. Most benchmarks have two types of question: multiple-choice question and open-ended question with numerical answer needed. We instantiate post prompt according to the question type:

> POST_PROMPT_NA: Please answer the question using a single word or phrase enclosed in backticks.
> POST_PROMPT_MCA: Answer with the option's letter from the given choices only, enclosed in backticks.

Given a user-prompt template, we construct the final user message by concatenating the user prompt with the question block, separated by blank lines and explicit field headers. For open-ended questions, we use:

> prompt = USER_PROMPT + "\n\nQuestion:\n" + question + "\n".

For multiple-choice questions, we additionally append the options block:

> prompt = USER_PROMPT + "\n\nQuestion:\n" + question + "\n\nOptions:\n" + options_str + "\n".

In all methods, the final answer must appear on the last line in the format Answer: 'X' to satisfy POST_PROMPT. (See more implementations in below scripts.)

A.2 User Prompts for Linguistic Reasoning Methods

We evaluate several advanced prompting strategies, including Chain-of-Thought (CoT) (Wei et al., 2022), Tree-of-Thoughts (ToT) (Yao et al., 2023), Least-to-Most (Zhou et al., 2023), and Cognitive Map prompting (Yang et al., 2025a). Below we provide the user-prompt templates used for each method.

Direct prompting. Direct prompting instructs the model to solve the task internally while suppressing any explicit reasoning, and to output only the final answer in the required format.

Direct Prompt

[System Prompt]
(Base system prompt; see §A.1.)

[User Prompt]
Reasoning protocol:
– Read the question carefully.
– You may think through the problem internally, but do *NOT* show your reasoning.
– Directly provide the final answer in the required format.

Output format:
Answer: 'X'

Now follow this protocol to answer the question below.

Chain-of-Thought (CoT) prompting. For CoT prompting (Wei et al., 2022), we explicitly request a step-by-step natural-language explanation before emitting the final answer line. The user prompt enforce a two-part output: a *Reasoning* block with the step-by-step explanation, followed by a *Final answer* block containing exactly Answer: 'X' on the last line.

Chain-of-Thought Prompt

[System Prompt]
(Base system prompt; see §A.1.)

[User Prompt]
Reasoning protocol:
– First, think step by step about the visual scene and the spatial relationships involved.
– Explain your reasoning clearly in natural language.
– At the end, provide a single final answer line in the required format.

Output format:
Reasoning:
[*step-by-step explanation*]
Final answer:
Answer: 'X'

Now follow this protocol to answer the question below.

Tree-of-Thoughts (ToT) prompting. For Tree-of-Thoughts (Yao et al., 2023), we enforces a three-stage procedure: (1) generate three distinct reasoning branches ("Thought 1–3"), (2) compare and

select the best branch under consistency and spatial-coherence checks, and (3) output the final answer based on the selected branch. The output format includes the three thoughts, an explicit evaluation section, the chosen best thought, and then the final answer line.

Tree-of-Thoughts Prompt

[System Prompt]
(Base system prompt; see §A.1.)

[User Prompt]
Reasoning protocol:
– **Step 1:** Generate multiple reasoning branches (*thoughts*).
– Propose 3 plausible reasoning paths about the question.
– Each path should be coherent and may use different assumptions about the spatial layout.
– **Step 2:** Evaluate and compare the thoughts.
– Check consistency with the video evidence, spatial coherence, and contradictions.
– Select the most reliable thought overall.
– **Step 3:** Produce the final answer using the best thought.
– Use the most reliable thought to derive a single final answer in the required format.

Output format:
Thought 1:
[*reasoning path 1*]
Thought 2:
[*reasoning path 2*]
Thought 3:
[*reasoning path 3*]
Evaluation:
– Thought 1: . . .
– Thought 2: . . .
– Thought 3: . . .
Best thought: Thought X because . . .
Final answer:
Answer: 'X'

Now follow this protocol to answer the question below.

Least-to-Most prompting. For Least-to-Most prompting (Zhou et al., 2023), we ask the model to decompose each question into ordered subproblems from easiest to hardest (e.g., object identification → local relations → global layout → final decision), solve them sequentially while reusing intermediate results, and finally provide a single answer line. The output format is constrained to three stages: decomposition, solving subproblems, and final answer.

Least-to-Most Prompt

[System Prompt]
(Base system prompt; see §A.1.)

[User Prompt]
Reasoning protocol:
– **Step 1:** Decompose the problem from easiest to hardest.
– Identify what type of visual-spatial task this is (e.g., object count, absolute distance,relative distance, relative direction, object size, room size, route planning, appearance order).
– Break the question into several subproblems ordered from the easiest to the most difficult.
– Typical stages may include:
– identifying relevant objects and regions,
– understanding local spatial relations,
– integrating them into a global spatial layout,
– making the final decision.
– **Step 2:** Solve subproblems in order.
– For each subproblem:
– Briefly name the subproblem.
– Explain how you solve it using evidence from the video.
– Reuse and refine results from previous subproblems.
– Keep the spatial layout consistent across all steps.
– **Step 3:** Produce the final answer.
– Based on the solved subproblems, give a single final answer in the required format.

Output format:
Step 1: Problem decomposition
– Subproblem 1: . . .
– Subproblem 2: . . .
– Subproblem 3: . . .
Step 2: Solving subproblems
– Subproblem 1:

– Reasoning: ...
– Subproblem 2:
– Reasoning: ...
– Subproblem 3:
– Reasoning: ...
Step 3: Final answer
Answer: 'X'

Now follow this protocol to answer the question below.

Cognitive Map prompting. For Cognitive Map prompting (Yang et al., 2025a), we use a specialized description prompt that asks the model to first produce a textual cognitive map: for a fixed set of indoor categories of interest (COI). For example, this is the categories for VSI-bench:

> ceiling light, trash can, bed, heater, closet, pillow, backpack, chair, refrigerator, tv, nightstand, keyboard, computer tower, coat hanger, table, trash bin, whiteboard, monitor, sofa, clock, computer mouse, radiator, telephone

the model estimates the center location of each object instance on a 10×10 grid and outputs a dictionary in strict JSON form:

> {"category name": ["(x_1, y_1)", ...], ...}.

After generating this cognitive map, the model answers the question, still respecting the same final answer constraint Answer: 'X'.

Cognitive Map Prompt

[System Prompt]
You are an expert specialized in solving spatial understanding questions using text descriptions of egocentric video sequences.
You are given:
– an egocentric video of an indoor environment,
– a question about that video.
You need to firstly generate a "Cognitive Map" derived from the egocentric video, and then answer the question as accurately as possible.

HOW TO GENERATE THE COGNITIVE MAP:
[Task]
This video captures an indoor scene. Your objective is to identify specific objects within the video, understand the spatial arrangement of the scene, and estimate the center point of each object, assuming the entire scene is represented by a 10x10 grid.
[Rule]
1. We provide the categories to care about in this scene: {CATEGORIES_OF_INTEREST}. Focus ONLY on these categories.
2. Estimate the center location of each instance within the provided categories, assuming the entire scene is represented by a 10x10 grid.
3. If a category contains multiple instances, include all of them.
4. Each object's estimated location should accurately reflect its real position in the scene, preserving the relative spatial relationships among all objects.

[Output]
Present the estimated center locations for each object as a list within a dictionary. *STRICTLY* follow this JSON format: {"CATEGORY NAME": ["(X_1, Y_1)", ...], ...}.

Answer format:
- {POST_PROMPT}
- Do NOT add extra text on the final answer line (no units, no explanations).

[User Prompt]
Reasoning protocol:
– Read the question and analyze the visual content carefully.
– Generate cognitive map as required to help determine spatial relationships and approximate distances.
– You may think through the problem internally, but do *not* show your reasoning.
– Directly provide the final answer in the required format.

Output format:
Final answer:
Answer: 'X'

Now follow this protocol to answer the question below.

Textual Representation of Allocentric Context from Egocentric Video prompting. We propose a new prompting method to elicit a struc-

tured intermediate representation before answering: the model first generates a structured Textual Representation of Allocentric Context from Egocentric Video that summarizes the room-aligned coordinate system, the camera trajectory, and an entity registry with timestamps and estimated positions. The textual representation must be a single YAML document with three top-level sections: `Meta_Context`, `Trajectory`, and `Entity_Registry`. Below is a detailed description which is included in the system prompt, and a reference TRACE prompt.

Meta_Context (required keys).

```
Meta_Context:
  room_topology: "<room shape/type>"
  grid_alignment:    "<what  +Y/+X  is
aligned with>"
  initial_camera_heading:     "<heading
relative to room grid>"
```

Trajectory (example).

```
Trajectory:
  # Track movement relative to the ROOM
GRID, not just camera view.
  - step: 0
    time: "0s"
    pos: [0.0, 0.0]
    facing: "NW (-X,+Y)"
    action:  "Standing  near  entrance,
panning across the room"
  - step: 1
    time: "4s"
    pos: [0.0, 1.8]
    facing: "North (+Y)"
    action:  "s forward along the main
room axis"
  - step: 2
    time: "8s"
    pos: [0.2, 3.5]
    facing: "East (+X)"
    action:  "Turning right to inspect
bedside area"
```

Entity_Registry (example).

```
  # The Map.  Coordinates are strictly
[x, y] in meters.
  - id: "door_01"
    category: "door"
    first_seen_at: "0s"
    state: "open"
    estimated_pos: [0.8, 0.0]
    approx_size: [0.9, 2.1, 0.1]
    visual_signature:    "White  hinged
door with silver handle"
    spatial_relation: "At the entrance
boundary of the bedroom"
  - id: "bed_01"
    category: "bed"
    first_seen_at: "5s"
    estimated_pos: [1.8, 2.8]
    approx_size: [1.6, 2.0, 0.6]
    orientation:  "Headboard against +X
wall"
    visual_signature:  "Double bed with
white sheets and dark frame"
    spatial_relation:    "Against   the
right wall, beside nightstand_01"
  - id: "nightstand_01"
    category: "nightstand"
    first_seen_at: "7s"
    estimated_pos: [1.9, 2.0]
    approx_size: [0.5, 0.6, 0.4]
    visual_signature:    "Small   wooden
bedside table"
    spatial_relation:    "In   front   of
bed_01 near the headboard"
  - id: "trash_bin_01"
    category: "trash_bin"
    first_seen_at: "10s"
    estimated_pos: [-1.3, 2.4]
    approx_size: [0.3, 0.4, 0.3]
    visual_signature:            "Black
cylindrical trash bin"
    spatial_relation:  "Near  the  left
wall below desk_01"
```

Textual Representation of Allocentric Context Prompt

[System Prompt]

You are an expert specialized in solving spatial understanding questions using text descriptions of egocentric video sequences.

You are given:

– an egocentric video of an indoor environment,

– a question about that video.

You need to firstly generate a structured "Textual Representation of Allocentric Context" derived from the egocentric video, and then answer the question as accurately as possible.

HOW TO GENERATE THE TEXTUAL REPRESENTATION OF ALLOCENTRIC CONTEXT:

1. Coordinate System Rules (Room-Aligned Allocentric Frame)

– Origin: the camera starting position is exactly `[0.0, 0.0]` on the floor plane.

– Major Axes (+Y / +X): Align the coordinate system with the *dominant walls* or *floor grid* of the room rather than the camera's initial viewing direction.:

– define '+Y' along that dominant structural direction;

– define '+X' as the perpendicular rightward direction in the floor plane.

– Units are approximate meters.

– Maintain one globally consistent coordi-

nate frame throughout the video. If the camera moves into another room, preserve the same global frame rather than resetting coordinates.

2. Meta-Context Rules
You must infer and report:
– `room_topology`: the overall spatial structure of the observed environment, such as 'rectangular bedroom', 'L-shaped office', or 'narrow hallway connected to kitchen'
– `grid_alignment`: the structural cue used to define the allocentric axes
– `initial_camera_heading`: the camera's initial facing direction relative to the room-aligned grid

3. Trajectory Rules
You must log the camera path continuously. Output a trajectory step for *every significant* camera movement.
– `step`: Sequential ID.
– `time`: Timestamp of the step (e.g., "2s") – `pos`: Estimated [x, y] of the camera.
– `facing`: Cardinal direction and axis (e.g., "North (+Y)").
– `action`: Short description of the camera motion or viewpoint change

4. Entity Registry Rules
You must register every visible entity individually. Never group objects. For each entity, include:
– `id`: unique identifier such as `chair_01`, `door_01`
– `category`
– `first_seen_at`
– `estimated_pos`: `[x, y]`
– `approx_size`: `[width, height, depth]`
– `visual_signature`: short appearance-based description for disambiguation
– `spatial_relation`: at least one relation to a nearby anchor or structure
– optional `state` when applicable (e.g., door open/closed, drawer open/closed)
– optional `orientation` when meaningful (e.g., bed headboard against east wall)

Output Format (Strict YAML)
refer to above

Hard Rules
– No omissions: if an object occupies more than 1% of pixels, it must be listed (including partial window/door edges).
– Force coordinates: you must estimate `[x,y]` for every listed item.
– Exhaustive count: never group items; if there are 6 chairs, I expect 6 entries in the registry.
– Global consistency: trajectory and entity coordinates must agree with one another.

Answer format:
- {POST_PROMPT}
- Do NOT add extra text on the final answer line (no units, no explanations).

[User Prompt]
Reasoning protocol:
– Read the question and analyze the visual content carefully.
– Generate the textual representation of allocentric context as required to determine spatial relations and approximate distances.
– You may think through the problem internally, but do *not* show your reasoning.
– Directly provide the final answer in the required format.

Output format:
Structured Textual Allocentric Representation:

Answer: 'X'

Now follow this protocol to answer the question below.

B Additional Results

B.1 Details on Decomposition Analysis

We present the per-category and per-task breakdowns of our compositional analysis in Tables 5 and 6, respectively. These results further reveal how different descriptor–parser combinations contribute to spatial reasoning performance and highlight key bottlenecks in modeling object-, relation-, and layout-level spatial concepts.

Table 5: *Decomposition analysis of 3D spatial QA performance (matrix).* We adopt the text-only prediction setting and report results for different combinations of visual descriptors and spatial knowledge parsers. Cell shows Avg. with (Multiple-Choice Answer / Numerical Answer).

Descriptor \ Parser	Gemini 3 Pro Avg. (MCA/NA)	Qwen2.5-72B-Instruct Avg. (MCA/NA)	Qwen2.5-32B-Instruct Avg. (MCA/NA)	Qwen2.5-7B-Instruct Avg. (MCA/NA)
Gemini 3 Pro	52.27 (44.12/57.40)	40.86 (42.33/39.47)	36.23 (38.03/41.35)	29.35 (30.00/28.73)
Qwen2.5-VL-72B-Instruct	36.11 (41.58/28.94)	31.11 (35.98/26.51)	26.12 (32.01/20.56)	24.06 (29.60/18.84)
Qwen2.5-VL-32B-Instruct	36.70 (40.74/32.18)	14.66 (29.20/0.94)	23.29 (29.60/17.34)	24.45 (30.24/18.99)
Qwen2.5-VL-7B-Instruct	32.72 (36.38/28.39)	29.08 (32.85/25.53)	24.33 (29.00/19.92)	25.19 (29.28/21.34)

Table 6: *Decomposition analysis of 3D spatial QA performance (grouped breakdown).* We adopt the text-only prediction setting and report results for different combinations of visual descriptors and spatial knowledge parsers.

Descriptor	Parser	Avg.	Numerical Answer				Multiple-Choice Answer			
			Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route	Order
Gemini 3 Pro	Gemini 3 Pro	52.27	28.52	32.66	67.28	48.02	49.58	62.66	52.43	64.93
	Qwen2.5-72B	40.86	28.12	28.91	55.38	39.69	39.30	40.60	34.54	50.97
	Qwen2.5-32B	36.23	27.45	21.55	49.81	44.10	40.00	30.99	30.41	45.15
	Qwen2.5-7B	29.35	24.62	22.95	36.13	29.03	28.87	28.31	26.29	35.11
Qwen2.5-VL-72B	Gemini 3 Pro	36.11	10.57	16.57	51.83	31.72	40.25	39.93	34.81	47.39
	Qwen2.5-72B	31.11	12.83	21.74	39.71	23.51	37.04	33.16	32.99	40.13
	Qwen2.5-32B	26.12	11.86	13.49	32.03	20.17	35.07	27.17	32.47	35.92
	Qwen2.5-7B	24.06	11.10	16.26	25.54	19.34	29.01	27.58	27.32	34.14
Qwen2.5-VL-32B	Gemini 3 Pro	36.70	14.93	19.07	52.20	38.20	36.93	39.58	34.27	48.78
	Qwen2.5-72B	14.66	0.00	2.87	0.10	0.00	35.21	30.68	33.51	19.42
	Qwen2.5-32B	23.29	0.00	13.45	36.26	0.00	29.86	40.70	31.96	11.49
	Qwen2.5-7B	24.45	0.00	27.75	20.45	26.08	29.15	38.74	33.51	17.15
Qwen2.5-VL-7B	Gemini 3 Pro	32.72	10.81	14.11	49.46	34.54	37.77	37.83	31.46	33.99
	Qwen2.5-72B	29.08	11.96	20.92	36.59	28.85	30.00	35.74	32.47	31.72
	Qwen2.5-32B	24.33	11.43	13.72	29.62	22.43	30.28	28.51	32.99	27.02
	Qwen2.5-7B	25.19	12.02	21.63	27.51	18.40	29.44	29.86	30.41	27.83

B.2 Instruction Following for Spatial Reasoning

In Tab. 7, we conduct additional experiments comparing Qwen-VL and Qwen LLM under the same textual representations. We find that Qwen-VL consistently underperforms the language-only model for both Gemini-generated and Qwen-generated representations. This observation suggests that visual instruction tuning can compromise spatial knowledge parsing ability, highlighting the importance of carefully designing visual training data to enable MLLMs to better capture 3D spatial concepts.

B.3 Detailed Results on Stratified Analysis

Tab. 8 provides a granular breakdown of model performance across the three distinct indoor scene datasets comprising VSI-Bench: ARKitScenes (Baruch et al., 2021), ScanNet (Dai et al., 2017), and ScanNetPP (Yeshwanth et al., 2023). Across both proprietary and open-weights architectures, our proposed TRACE prompting robustly yields performance improvements over the Direct baseline within each environment distribution. Notably, TRACE achieves balanced gains without overfitting to a specific dataset's spatial characteristics, confirming the reliable cross-environment generalization of our textual allocentric representation.

B.4 Full Evaluation Results on VSI-Bench

The complete results for VSI-Bench are detailed in Table 9. To mitigate the risk of data contamination, we restrict our evaluation to model versions released no later than six months after the publication of VSI-Bench and OST-Bench. Consequently, our final selection includes Gemini 3 Pro (Gemini Team, 2025), o3 (OpenAI, 2025), Qwen2.5-VL (Yang et al., 2024), MiMo-VL-7B (Xiaomi, 2025), and GLM-4.5V (V Team et al., 2025).

B.5 Token Efficiency

As shown in the Tab. 10, the token consumption of TRACE varies depending on the underlying model's generation tendencies but generally maintains a highly favorable performance-to-cost

Table 7: Effect of visual tokens and multimodal training on spatial context reasoning.

Descriptor	Parser	Avg.	Numerical Answer				Multiple-Choice Answer			
			Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route	Order
Gemini 3 Pro	Qwen2.5-72B	40.86	28.12	28.91	55.38	39.69	39.30	40.60	34.54	50.97
Gemini 3 Pro	Qwen2.5-VL-72B	39.48	28.21	20.56	54.23	42.15	44.37	34.50	39.18	53.56
Qwen2.5-VL-72B	Qwen2.5-72B	31.11	12.83	21.74	39.71	23.51	37.04	33.16	32.99	40.13
Qwen2.5-VL-72B	Qwen2.5-VL-72B	26.65	9.08	11.08	36.21	20.59	28.31	30.58	36.08	40.78

trade-off, particularly when compared to highly branching reasoning methods. For instance, on the compact MiMo-VL-7B model, TRACE consumes significantly fewer tokens (737.72) than Tree-of-Thoughts (1132.86) and Least-to-Most (1097.05) prompting, while simultaneously delivering superior average performance. However, for several other large foundation models, including Gemini 3 Pro, o3, and Qwen2.5-VL-72B, our method is noticeably more token-intensive than these baselines. This increased consumption is an expected trade-off, as explicitly generating a structured allocentric representation inherently loads the context window with an exhaustive spatial cache. While the consistent accuracy gains across diverse models justify this computational overhead, optimizing token efficiency during the structured reasoning process constitutes a largely orthogonal research direction, which we leave for future work.

Table 8: *Stratified VSI-Bench results across underlying scene datasets.* Performance is broken down by ARKitScenes, ScanNet, and ScanNetPP for both proprietary and open-weight models. TRACE consistently improves over the Direct baseline across all environment distributions, indicating strong cross-environment generalization.

Method	Avg.	ARKitScenes	ScanNet	ScanNetPP
o3 as base model				
Direct	51.15	49.28	53.55	49.78
CoT	52.36	51.76	53.53	51.36
ToT	52.09	51.27	53.46	51.06
LtM	52.50	51.70	54.31	50.80
CM	53.93	52.69	**56.24**	52.00
Ours	**54.08**	**54.63**	55.06	**52.11**
MiMo-VL-7B-SFT as base model				
Direct	39.79	39.36	39.97	**40.01**
CoT	37.49	37.33	37.08	38.26
ToT	39.14	38.45	40.50	37.97
LtM	38.34	36.52	39.52	38.66
CM	36.85	36.31	36.99	37.23
Ours	**41.42**	**42.47**	**41.72**	39.82
Qwen2.5-VL-72B-Instruct as base model				
Direct	36.28	36.05	36.20	36.65
CoT	29.78	27.76	31.37	29.74
ToT	38.06	40.43	36.84	37.19
LtM	38.01	40.41	36.74	37.18
CM	35.47	37.83	35.07	33.45
Ours	**39.38**	**42.03**	**37.80**	**38.73**
GLM-4.5V as base model				
Direct	37.33	33.97	38.58	39.26
CoT	38.48	36.58	39.81	38.73
ToT	40.66	39.07	41.44	41.30
LtM	40.99	39.16	42.13	41.44
CM	38.93	37.92	39.03	39.89
Ours	**45.01**	**46.83**	**44.37**	**43.89**
Gemini 3 Pro as base model				
Direct	52.61	52.26	52.61	52.99
CoT	53.65	52.45	53.68	54.93
ToT	58.88	57.61	58.32	**61.10**
LtM	59.52	58.01	59.78	60.83
CM	59.72	**59.50**	59.54	60.23
Ours	**60.15**	59.39	**60.42**	60.63

Table 9: *Evaluation results on the VSI benchmark.* We report average performance and detailed breakdowns across numerical-answer and multiple-choice tasks, under proprietary and open-sourced base models. Best results are in **bold**, and second-best are underlined.

Methods	Avg.	Numerical Answer				Multiple-Choice Answer			
		Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route	Order
Gemini 3 Pro as base model									
Direct	52.61	33.77	32.57	67.09	42.99	62.54	50.52	51.03	70.71
CoT	53.65	30.35	34.54	64.05	40.76	61.78	58.09	**61.34**	71.96
ToT	58.88	44.55	**42.12**	72.20	45.55	65.35	57.83	55.62	**73.73**
LtM	59.52	45.19	40.72	73.36	44.15	**65.82**	60.40	53.59	73.64
CM	59.72	46.70	41.43	72.49	**50.14**	63.69	58.62	55.50	72.61
Ours	**60.15**	**47.55**	38.82	**73.90**	45.62	63.85	**61.70**	58.01	72.97
Qwen2.5-VL-72B-Instruct as base model									
Direct	36.28	**33.36**	20.53	49.31	41.49	**43.38**	27.79	32.47	**44.01**
CoT	29.78	21.27	24.95	16.31	40.94	39.44	33.16	28.87	43.53
ToT	38.06	17.89	26.20	53.15	**47.01**	41.55	36.78	**35.05**	**44.01**
LtM	38.01	23.27	**31.39**	54.49	38.68	42.96	34.71	29.90	36.73
CM	35.47	21.58	15.67	52.65	37.26	39.44	36.05	34.54	42.39
Ours	**39.38**	22.05	28.03	**59.98**	38.99	40.85	**37.40**	31.96	42.56
MiMo-VL-7B as base model									
Direct	39.79	**36.02**	29.84	52.38	42.95	40.14	33.78	31.44	47.41
CoT	37.49	34.27	23.50	48.52	43.23	38.73	32.75	27.84	49.23
ToT	39.14	29.45	30.44	54.26	40.14	41.41	32.02	32.47	46.60
LtM	38.34	35.09	24.47	48.22	**44.48**	**43.10**	30.79	**35.05**	49.50
CM	36.85	27.43	23.14	50.14	39.06	41.41	32.54	27.84	46.76
Ours	**41.42**	33.27	**31.51**	**58.67**	41.56	39.44	**35.33**	28.87	**51.29**
o3 as base model									
Direct	51.15	33.26	31.95	69.37	52.57	58.87	44.11	42.78	69.42
CoT	52.36	34.11	28.37	69.81	50.31	**59.72**	48.89	57.06	70.96
ToT	52.09	40.07	24.26	69.55	48.68	59.15	50.23	55.35	69.36
LtM	52.50	35.68	26.98	70.05	47.05	59.15	50.97	57.96	71.22
CM	53.93	34.18	**33.35**	70.19	52.05	59.30	**51.10**	**62.99**	**71.26**
Ours	**54.08**	**43.40**	29.93	**72.48**	**54.03**	57.32	49.83	56.10	70.02
GLM-4.5V as base model									
Direct	37.33	34.87	32.74	28.13	29.72	47.32	39.05	35.57	49.92
CoT	38.48	33.42	31.07	39.88	25.52	45.49	39.26	35.57	50.19
ToT	40.66	34.45	32.17	47.29	27.81	45.63	**39.77**	32.47	51.88
LtM	41.35	32.32	**33.14**	51.26	22.26	**49.30**	36.47	**39.18**	**58.00**
CM	38.93	37.77	31.91	36.61	30.07	45.21	37.40	33.51	54.05
Ours	**45.01**	**40.41**	32.84	**65.11**	**36.74**	45.77	38.53	36.60	50.68

Table 10: *Token consumption and performance trade-offs of prompting methods across different models.* TRACE generally maintains a favorable performance-to-cost ratio relative to branching reasoning baselines, although its token usage varies by backbone. Best results for each model are in **bold**.

	GLM-4.5V		MiMo-VL-7B		Qwen2.5-VL-72B		o3		Gemini 3 Pro	
Method	Tok	Avg	Tok	Avg	Tok	Avg	Tok	Avg	Tok	Avg
Direct	405.17	37.33	337.36	39.79	3.48	36.28	3.61	51.15	334.35	52.61
CoT	568.36	38.48	579.21	37.49	129.62	28.44	76.20	52.36	479.64	53.65
ToT	1079.97	40.66	1132.86	39.14	308.99	37.68	352.30	52.09	450.82	58.88
LtM	989.55	40.99	1097.05	38.34	229.84	38.01	220.28	52.50	571.88	59.52
CM	722.16	38.93	723.68	36.85	224.72	35.47	81.07	53.93	403.23	59.72
Ours	967.91	**45.01**	737.72	**41.42**	755.87	**39.38**	435.49	**54.08**	843.91	**60.15**