SpatiO: Adaptive Test-Time Orchestration of Vision-Language Agents for Spatial Reasoning

Chan Yeong Hwang[1], Miso Choi[1], Sunghyun On[3], Jinkyu Kim[1,2], Jungbeom Lee[1†]

[1]Korea University [2]Kakao Mobility Corp. [3]Handong Global University
South Korea
{flaxinger, miso8070, jinkyukim, jbeomlee}@korea.ac.kr
22100457@handong.ac.kr

[†]Co-corresponding authors.

Abstract. Understanding visual scenes requires not only recognizing objects but also reasoning about their spatial relationships. Unlike general vision-language tasks, spatial reasoning requires integrating multiple inductive biases, such as 2D appearance cues, depth signals, and geometric constraints, whose reliability varies across contexts. This motivates *spatial adaptability*: the ability to flexibly coordinate reasoning strategies depending on the input. However, most existing approaches rely on a single reasoning pipeline that implicitly learns a fixed spatial prior, limiting their ability to adapt under distribution changes. Multi-agent systems offer a promising alternative by aggregating diverse reasoning trajectories, but prior spatial reasoning attempts primarily employ homogeneous agents, restricting the inductive-bias diversity they can leverage. In this work, we introduce **SpatiO**, a heterogeneous multi-agent framework for spatial reasoning that coordinates multiple vision-language specialists with complementary inductive biases. To enable effective collaboration, we propose **Test-Time Orchestration (TTO)**, a calibration mechanism that dynamically evaluates and reweights agents based on their observed reliability during inference, without modifying model parameters. Extensive experiments on diverse spatial reasoning benchmarks, including 3DSRBench, STVQA-7k, CV-Bench, and Omni3D-Bench, demonstrate that SpatiO consistently improves spatial reasoning performance over both closed-source and open-source baselines. The project page is available at https://cy-h1329.github.io/spatio/.

Keywords: Spatial Reasoning · Multi-Agent Systems · Vision-Language Models · Test-Time Orchestration

1 Introduction

Understanding visual scenes requires more than recognizing objects and attributes; it demands reasoning about how entities are arranged in space, how they relate to one another, and how these relationships change across viewpoints [19, 26, 29, 31]. This capability, commonly referred to as **spatial reasoning**, is fundamental for visual intelligence. It is becoming increasingly essential

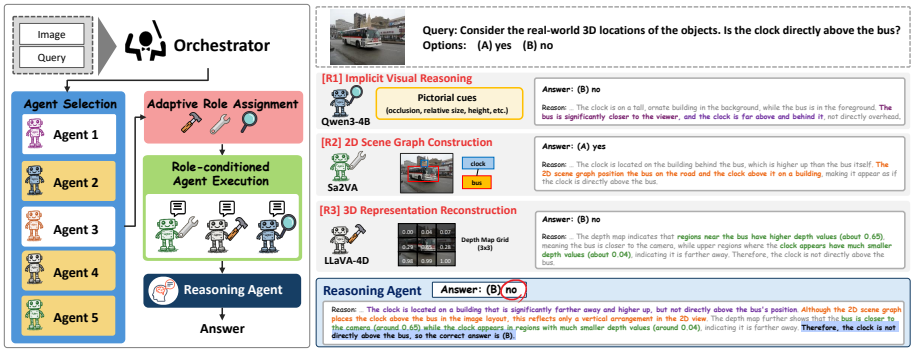


Fig. 1: Overall framework illustration of SpatiO. Multiple heterogeneous vision-language agents operate as specialists under different roles, and their outputs are aggregated to produce the final spatial reasoning result.

as artificial systems aim to interpret, navigate, and interact with real-world environments, ranging from embodied agents, robotics, to augmented and virtual reality, 3D scene understanding, and physical world modeling. Unlike general visual tasks such as image captioning or visual question answering, spatial reasoning requires integrating heterogeneous sources of information, such as 2D appearance cues, depth signals, spatial relations, and implicit geometric priors [36]. Crucially, the reliability of these cues is highly context-dependent. In canonical views, appearance cues are often sufficient, whereas geometric constraints become critical under occlusion or clutter. As a result, effective performance of spatial reasoning hinges on **spatial adaptability**, the ability to flexibly coordinate different strategies in response to the specific conditions of a given task.

The predominant approach for spatial reasoning has been large-scale supervision: training on billions of synthetic spatial VQA examples [7], or incorporating structured representations of location and orientation derived from 3D data [21]. These models achieve competitive performance by internalizing powerful spatial priors tailored to their training distributions, effectively functioning as specialists for particular categories of spatial queries. Subsequent approaches [4, 20] further enhance reasoning by introducing explicit structural priors and reinforcement learning, improving data efficiency and inference quality within a single reasoning pipeline. Despite these advances, existing approaches share a common limitation: they realize spatial reasoning through a single-model reasoning pipeline that implicitly internalizes a fixed spatial prior tied to its training distribution. In practice, this induces strong but narrow inductive biases: the model behaves as a specialist for the configurations and query types it has frequently seen, while generalizing poorly to out-of-distribution layouts or viewpoints [19, 21]. Since spatial reasoning is inherently multifaceted, spanning depth estimation, orientation inference, and multi-object relational reasoning, any single model tuned to a fixed spatial prior inevitably overfits certain facets of this space while neglecting others. As also observed in recent spatial benchmarks, different models excel on different spatial categories, with decorrelated failure modes across depth, orientation, and relational tasks [19,23]. Even methods that introduce in-

termediate reasoning steps remain confined to a fixed architectural perspective, limiting their ability to integrate complementary spatial cues from structurally diverse models.

However, spatial reasoning fundamentally involves multiple inductive biases, whose reliability is inherently context-dependent rather than fixed. This raises a fundamental question:

> How can a system dynamically determine which reasoning strategies to trust for a given input and modulate their contributions accordingly at test time?

With this perspective, we reframe spatial reasoning as a coordination problem, where multiple complementary reasoning strategies are adaptively combined at test time rather than optimized within a single static model.

Prior work [30, 32] has shown that heterogeneous agents with decorrelated failure modes improve robustness under distribution shift, suggesting that separating inductive biases across specialists offers a principled path toward adaptability. However, existing attempt [22] to apply this heterogeneous-agent paradigm to spatial reasoning remain limited: replicating a homogeneous backbone or restricting agents to peripheral roles fails to address conditional reliability across reasoning strategies. As a result, the system neither maintains role-dependent estimates of when a given strategy is trustworthy, nor down-weights specialists when their outputs are noisy. Because multi-step pipelines amplify upstream errors, effective orchestration must dynamically reweight contributions based on context-specific reliability, motivating a framework for reliability-aware collaboration among complementary VLM specialists.

We introduce **SpatiO**, a heterogeneous multi-agent framework for spatial reasoning through adaptive test-time coordination. SpatiO assembles a diverse pool of VLMs with distinct architectures, training objectives, and geometric inductive biases, each independently solving the spatial query under a designated reasoning role. We propose a novel **Test-Time Orchestration (TTO)** mechanism that dynamically reweights agents based on observed reliability, leveraging per-agent confidence scores to continuously update the orchestration strategy, enabling parameter-free adaptation without modifying model weights. By adapting collaboration rather than parameters, SpatiO avoids catastrophic forgetting, incurs minimal overhead, and remains compatible with both open-source and black-box models. Our primary contributions are as follows:

1. We propose **SpatiO**, a fully heterogeneous, role-based multi-agent framework for spatial reasoning that coordinates complementary pretrained VLMs without any parameter updates.
2. We formulate **Test-Time Orchestration (TTO)**, a parameter-free mechanism that dynamically adjusts agent selection and aggregation weights via Bayesian trust modeling and dual EMA filtering.
3. We empirically show that SpatiO outperforms monolithic VLMs across challenging spatial benchmarks including 3DSRBench, STVQA, CV-Bench and Omni3D-Bench.

2 Related Work

2.1 Spatial Intelligence in Vision-Language Models

Research on enhancing spatial reasoning in VLMs has progressed from implicit feature learning to explicit geometric modeling. SpatialVLM [7] trained on billions of synthetic spatial VQA pairs to facilitate quantitative distance estimation, while SpatialLLM [21] transformed 3D spatial data into structured textual representations. Although these data-intensive methods demonstrate strong in-distribution performance, models such as SpatialLadder [17] still exhibit noticeable degradation when task categories or data distributions differ from training, arising from reliance on distribution-specific spatial heuristics rather than compositional reasoning.

Later research introduced explicit structural priors to address these limitations. SpatialRGPT [9] integrated 3D region proposals and scene-graph priors directly into VLM attention mechanisms. SpatialThinker [4] combined scene-graph grounding with online reinforcement learning guided by a multi-objective reward, achieving state-of-the-art results with only 7K training samples. Spatial-Reasoner-R1 [20] utilized fine-grained direct preference optimization (fDPO) and long chain-of-thought supervision to systematically parse complex spatial configurations. Orthogonal to this, RegionVLM [16] adds explicit regional grounding without architectural changes, a further inductive bias for pools such as ours. However, all of these approaches realize spatial reasoning through a single, fixed inference pipeline, limiting adaptation to out-of-distribution configurations or novel viewpoints.

2.2 Multi-Agent Systems for Spatial Reasoning

The ensembling of multiple LLMs has demonstrated significant effectiveness in language reasoning tasks. MoA [30] and MoSA [32] show that synthesizing diverse reasoning trajectories from heterogeneous models consistently outperforms individual model performance. Notably, *Debate or Vote* [10] establishes that when agents share the same underlying model, inter-agent debate does not improve expected correctness, underscoring that the benefits of multi-agent systems are fundamentally dependent on genuine architectural heterogeneity.

In the visual domain, VADAR [22] outperforms program-synthesis baselines [13,28] but depends on a single GPT-4o backbone with rule-based aggregation. This approach exemplifies the homogeneous failure mode described in *Debate or Vote*, resulting in underperformance compared to a standard single-pass GPT-4o baseline. SpatiO addresses these limitations by pooling vision-language models (VLMs) with decorrelated error distributions and replacing rule-based aggregation with trust-weighted, geometry-aware re-generation.

More recently, GCA [8] assigns dual roles within a single Qwen3-VL-Thinking backbone, improving spatial inference robustness within a unified system, but remains unable to leverage complementary inductive biases across heterogeneous specialists.

2.3 Reliability Estimation and Test-Time Adaptation

A primary challenge in multi-agent spatial reasoning is estimating agent reliability under dynamic conditions. MATE [1] addresses this via a Beta–Bernoulli trust model updated online, but is designed for abstract cooperative agents without structured geometric evidence or role-dependent behaviors in VLMs.

At a finer granularity, [11] shows that context-truthful attention heads, inherited within model lineages, can be selectively amplified to improve contextual grounding, complementary to SpatiO's agent-level trust estimation.

Test-Time Adaptation (TTA) [18, 27] adapts models via parameter-level updates (*e.g.*, entropy minimization), but is restricted to single-model backbones with mutable weights, limiting applicability when coordinating black-box VLMs.

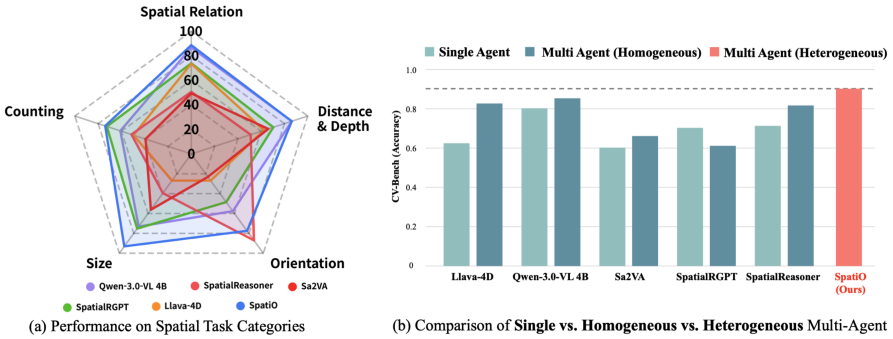
SpatiO integrates both paradigms into a spatially grounded formulation: rather than adapting parameters, test-time adaptation operates at the orchestration level by maintaining a Bayesian trust estimator over agent–role–category triplets, smoothed via dual EMA and used to dynamically reweight heterogeneous specialists. This yields a reliability-aware coordination mechanism tailored to multimodal spatial reasoning, rather than generic agent interaction or single-model adaptation.

3 Observation

3.1 Decorrelated Error Patterns across Spatial Specialists

To better understand the limitations of existing spatial reasoning systems, we analyze five representative vision-language models spanning diverse training paradigms, including large-scale synthetic spatial VQA supervision, structured preference alignment, and reinforcement learning with verifiable rewards. The models further differ in their representational assumptions, ranging from dense 2D grounding mechanisms to explicit 3D region-aware attention and scene-graph-based reasoning. We evaluate these models on 3DSRBench [19], STVQA-7k [4], and CV-Bench [29], reorganizing task categories into five groups: Spatial Relation, Counting, Size, Distance & Depth, and Orientation.

Fig. 2.(a) presents a radar plot summarizing per-group performance across models. A consistent pattern emerges: no single model uniformly excels across all spatial groups. Instead, each model exhibits category-specific strengths accompanied by systematic weaknesses. For instance, SpatialReasoner [20], which incorporates explicit multi-stage 3D structural representations, performs strongly on orientation-sensitive queries but shows degradation on distance and depth estimation tasks. In contrast, SpatialRGPT [9], which leverages 3D scene graphs for data generation and integrates depth-aware features, demonstrates stronger performance on distance and depth–related queries while underperforming on other relational categories.



(a) Performance on Spatial Task Categories

(b) Comparison of **Single vs. Homogeneous vs. Heterogeneous** Multi-Agent

Fig. 2: (a) Per-spatial task accuracy of five spatial reasoning specialists (single-agent) and SPATIO (Ours) on the reorganized task groups. Each agent occupies a distinct region of the performance space, motivating heterogeneous orchestration over single-backbone selection.

3.2 Homogeneous vs. Heterogeneous Multi-Agent Systems

Building on the specialist behaviors identified in Section 3.1, a logical progression is to deploy multiple agents and aggregate their outputs, given that no single model is universally reliable. Prior work such as VADAR [22] explores homogeneous multi-agent systems for spatial reasoning, deploying multiple instances of a single backbone. However, as shown by [10], such ensembles yield only marginal gains: agents sharing the same architecture exhibit identical inductive biases and failure modes, causing aggregation to amplify shared errors rather than correct them. To verify this (Fig. 2b), we deploy three instances of the same backbone with distinct role-based prompts. While aggregation improves over the single-agent baseline, confirming that prompt-level diversity introduces variation in reasoning trajectories, the gains remain modest due to shared architecture and training-induced biases. This motivates **architectural heterogeneity**. Combining distinct spatial specialists (e.g., Qwen3-VL-4B and SpatialReasoner) yields substantially larger gains across benchmarks, as each model contributes complementary geometric perspectives inaccessible within any single backbone.

4 Method

4.1 Overall Framework

As established in Section 1, no single VLM demonstrates uniform superiority across spatial task categories, and their error distributions are largely uncorrelated (Fig. 2). SPATIO leverages this complementarity by treating pre-trained VLMs as fixed specialist members within a structured team, replacing parameter-level adaptation with orchestration-level optimization. Fig. 3 illustrates the overall pipeline, which comprises three sequential components: *Query Routing*, *Role-Based Analysis*, and *Weighted Synthesis* (Section 4.3).

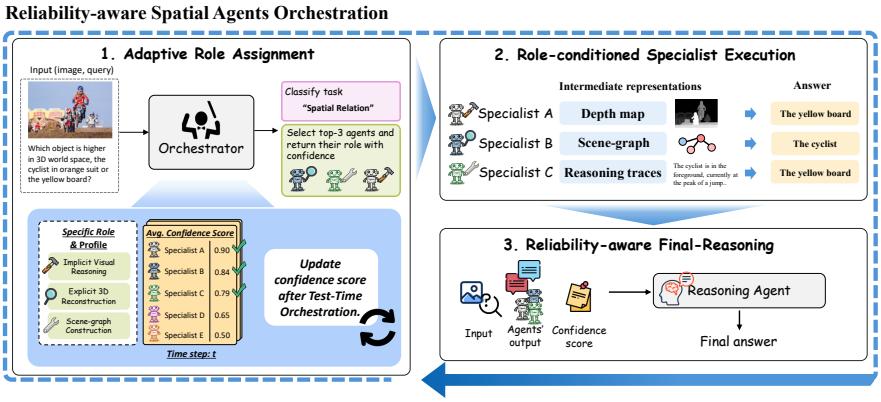


Fig. 3: Overview of the SpatiO framework. *(1)* The head agent classifies the query and selects the Top-3 agents with role assignments and trust weights. *(2)* Each agent independently reasons under its designated role, optionally invoking open-source tools; outputs are stored in shared memory $\mathcal{M}^{(t)}$. *(3)* The reasoning agent performs conditional evidence integration and emits $\hat{y}^{(t)}$; trust scores are updated via Bayesian estimation and dual EMA.

4.2 Preliminaries

Agent Pool. Our candidate pool $\mathcal{A}$ consists of five open-source vision-language models (VLMs) selected to maximize architectural complementarity. All models can be deployed directly at test time without the need for additional fine-tuning.

Role Definitions. Each agent in $\mathcal{A}$ can be assigned one of three canonical roles $k \in \mathcal{K}$ via structured system prompts specifying the reasoning strategy, tools, and output format, *prompt-mediated role injection*, enabling any model to assume any role without parameter modification.

1. **Implicit Visual Reasoning**: Performs end-to-end spatial visual question answering (VQA) using the raw 2D image, relying solely on the agent's inherent visual grounding capabilities without external tools.
2. **Explicit 3D Reconstruction**: Utilizes DepthPro [5] for metric monocular depth estimation and SAM2 [25] for instance segmentation, integrating these outputs into a 3D point-cloud representation before generating an answer.
3. **Scene-Graph Construction**: Constructs a relational scene graph using DINO-v2 [24], encoding object identities and spatial predicates such as *above*, *left-of*, and *closer-than*. The agent then grounds its answer on this symbolic structure [9, 14].

4.3 Spatial Multi-agentic Orchestration

Adaptive Role Assignment. The head agent classifies the incoming pair $(x^{(t)}, I^{(t)})$, consisting of a textual query $x^{(t)}$ and an observed image $I^{(t)}$, into one

i: agent, k: role, c: category, t: time step.

Symbol	Meaning
$s_{i,k,c}^{(t)}$	Confidence score of agent i *under role* k for category c
$\bar{s}_{i,c}^{(t)}$	Role-averaged confidence of agent i for category c; used for Top-3 *agent* selection
$w_{i,k,c}^{(t)}$	Normalized reliability *weight* of agent i in role k for category c; used in final reasoning

of $|\mathcal{C}|$ spatial task categories from a unified taxonomy derived from Section 3.1, where $\mathcal{C} = \{$*Spatial Relation, Counting, Size, Distance & Depth, Orientation*$\}$ with $|\mathcal{C}| = 5$. t denotes the orchestration time step (Section 4.4). For each agent i, role $k \in \mathcal{K}$, and category c, we maintain a confidence score $s_{i,k,c}^{(t)}$, initialized to 0.5 for unseen categories, and select the top-3 agents via an agent-level aggregated score:

$$\bar{s}_{i,c}^{(t)} = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} s_{i,k,c}^{(t)}, \tag{1}$$

$$\mathcal{S}_c^{(t)} = \mathrm{Topk}_i \left(\bar{s}_{i,c}^{(t)},\, k = 3 \right), \quad \bar{s}_{i,c}^{(t)} \in [0, 1], \tag{2}$$

where the selection is performed over all agents $i \in \mathcal{A}$ for the current query category c. For each predefined role $k \in \mathcal{K}$, we assign the agent within $\mathcal{S}_c^{(t)}$ that achieves the highest confidence score $s_{i,k,c}^{(t)}$ (ties are resolved using $\bar{s}_{i,c}^{(t)}$). We denote by k_i the role assigned to agent i under this selection rule. To quantify role-specific reliability, we compute normalized weights:

$$w_{i,k,c}^{(t)} = \frac{\exp(\beta\, s_{i,k,c}^{(t)})}{\sum_{k'} \exp(\beta\, s_{i,k',c}^{(t)})}, \tag{3}$$

where β controls the sharpness of the assignment. These weights encode the relative reliability of assigning role k to agent i for category c, and are later used as conditioning signals in the reliability-aware final reasoning stage (Section 4.3). This stage outputs the routing plan $\{(i,\, k_i,\, w_{i,k_i,c}^{(t)})\}_{i \in \mathcal{S}_c^{(t)}}$, which specifies, for the current query, the selected agents, their assigned roles, and their role-conditioned reliability weights.

Role-conditioned Specialist Execution. For each query, each selected agent processes $(x^{(t)}, I^{(t)})$ independently under its designated role prompt, producing a primary answer $a_i^{(t)}$ and intermediate representations $z_i^{(t)}$ (e.g., depth maps, scene-graph structures, or reasoning traces). All outputs are written to the shared evidence pool $\mathcal{E}^{(t)}$, preserving the full evidential chain (e.g., CoT), not only final predictions, for subsequent reliability-aware final reasoning within the same step.

Reliability-aware Final Reasoning. For final synthesis, let F denote the reasoning agent (DeepSeek-VL-R1-7B). Given the input pair $(x^{(t)}, I^{(t)})$ and the routing plan, the final prediction is produced as:

$$\hat{y}^{(t)} = F\Big(x^{(t)},\, I^{(t)},\, \mathcal{E}^{(t)},\, \{(a_i^{(t)}, k_i,\, w_{i,k_i,c}^{(t)})\}_{i \in \mathcal{S}^{(t)}} \Big), \tag{4}$$

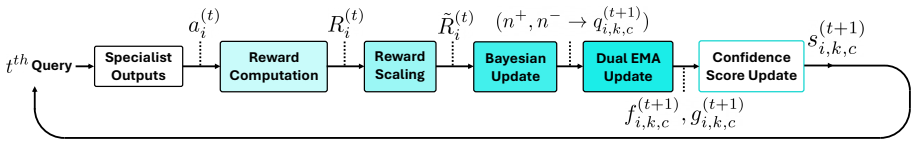


Fig. 4: Overall framework of the proposed Test-time Orchestration (TTO).

where $\mathcal{E}^{(t)}$ contains all primary answers and intermediate outputs generated by the selected specialists at step t. Rather than performing simple weighted voting, F conducts conditional re-generation by jointly considering agreement and divergence patterns across specialist outputs. The intermediate representations stored in $\mathcal{E}^{(t)}$ provide structured spatial cues beyond final answers of specialists. The weights $w_{i,k_i,c}^{(t)}$ modulate the influence of each role during final reasoning, prioritizing agents with higher estimated reliability weights for the current task category. After producing $\hat{y}^{(t)}$, the confidence score update described in Section 4.4 is triggered.

4.4 Test-Time Orchestration

Formulation. For each triple (i, k, c), we maintain a confidence score $s_{i,k,c}^{(t)} \in [0, 1]$ quantifying the expected reliability of agent i under role k on task category c at time t. The Test-Time Orchestration (TTO) objective is to improve the final answer accuracy over the query stream by updating these scores from observed outcomes, with zero parameter updates to any model.

Reward Computation. We distinguish two supervision regimes based on whether the final reasoning output $\hat{y}^{(t)}$ matches the ground truth $y^{(t)}$ (agreement) or not (divergence). Importantly, similarity is always computed at the specialist level by comparing each agent's output $a_i^{(t)}$ to the ground truth, regardless of the correctness of the final reasoner.

Here, $\mathrm{sim}(\cdot, \cdot)$ denotes a normalized semantic similarity between two textual answers. In our implementation, we compute the cosine similarity between their sentence embeddings and rescale it to $[0, 1]$:

$$\mathrm{sim}(u, v) = \frac{1}{2}\left(\frac{\phi(u)^\top \phi(v)}{\|\phi(u)\|_2 \|\phi(v)\|_2} + 1\right), \tag{5}$$

where $\phi(\cdot)$ denotes a sentence embedding encoder. For multiple-choice benchmarks, answers are first normalized to their canonical option text before computing similarity, so that semantically equivalent outputs such as "(A) yes" and "yes" are treated consistently.

A soft reward is defined as:

$$R_i^{(t)} = \begin{cases} 2 \cdot \mathrm{sim}(a_i^{(t)}, y^{(t)}) - 1 & \hat{y}^{(t)} = y^{(t)}, \\ 2 \cdot \mathrm{sim}(a_i^{(t)}, y^{(t)}) - 1 - \kappa\, \delta_i^{(t)} & \hat{y}^{(t)} \neq y^{(t)}, \end{cases} \tag{6}$$

where $\delta_i^{(t)}$ denotes a relative underperformance penalty applied in the divergence

case. It is defined as

$$\delta_i^{(t)} = \max\big(0,\, \mathrm{sim}(\hat{y}^{(t)}, y^{(t)}) - \mathrm{sim}(a_i^{(t)}, y^{(t)})\big). \tag{7}$$

Even when the final reasoning is correct, individual specialists may still be partially incorrect; conversely, under divergence, some specialists may have produced answers closer to the ground truth than the final output. Thus, the supervision signal for each specialist is defined at the individual level, rather than being determined by the correctness of the final reasoning outcome. In the divergence case, $\delta_i^{(t)}$ additionally penalizes agents whose outputs are less similar to the ground truth than the already incorrect final reasoning output [12]. Thus, agents that underperform relative to the failed consensus incur a larger penalty, controlled by $\kappa > 0$.

To prevent premature trust divergence in sparsely observed categories, the reward is further scaled by a category-dependent ramp factor:

$$\tilde{R}_i^{(t)} = \phi(N_c) \cdot R_i^{(t)}, \qquad \phi(N_c) = 1 - \exp\big(-\tfrac{N_c}{T}\big)\,, \tag{8}$$

where N_c is the cumulative query count for category c, and $T > 0$ is a ramp temperature controlling how quickly full trust adaptation activates (smaller T: faster ramp-up). In our experiments, we set $T = 5$ unless otherwise specified.

Bayesian Trust Update. To accumulate long-term reliability statistics for each (i, k, c) triple, we maintain a Beta-Bernoulli trust model [15]. Here, $n^+_{i,k,c}(t)$ and $n^-_{i,k,c}(t)$ denote the cumulative soft counts of positive and negative evidence, respectively, for agent i under role k in category c up to step t.

The scaled reward $\tilde{R}_i^{(t)} \in [-1, 1]$ is first mapped to $[0, 1]$ via $\tilde{r}^{(t)}_{i,k,c} = (\tilde{R}_i^{(t)} + 1)/2$, which can be interpreted as a fractional success probability. The sufficient statistics are then updated as:

$$n^+_{i,k,c}(t{+}1) = n^+_{i,k,c}(t) + \tilde{r}^{(t)}_{i,k,c}, \tag{9}$$

$$n^-_{i,k,c}(t{+}1) = n^-_{i,k,c}(t) + (1 - \tilde{r}^{(t)}_{i,k,c}). \tag{10}$$

Intuitively, high-reward steps contribute more mass to the positive count, while low-reward steps increase the negative count, enabling gradual and noise-tolerant confidence score adaptation over time. The posterior mean is computed as:

$$q^{(t+1)}_{i,k,c} = \frac{n^+_{i,k,c}(t+1)}{n^+_{i,k,c}(t+1) + n^-_{i,k,c}(t+1)}. \tag{11}$$

This quantity serves as a calibrated estimate of long-run reliability.

Dual Exponential Moving Average. To balance rapid adaptation with long-term stability, we maintain two exponential moving averages (EMA) [12] for each (i, k, c) triple:

$$f^{(t+1)}_{i,k,c} = (1{-}\lambda_f)\, f^{(t)}_{i,k,c} + \lambda_f\, \tilde{R}_i^{(t)}, \tag{12}$$

$$g^{(t+1)}_{i,k,c} = (1{-}\lambda_g)\, g^{(t)}_{i,k,c} + \lambda_g\, q^{(t+1)}_{i,k,c}, \tag{13}$$

The short-term component $f_{i,k,c}$ tracks recent reward signals $\tilde{R}_i^{(t)}$, allowing the system to quickly react to sudden performance changes. In contrast, the long-term component $g_{i,k,c}$ smoothly accumulates the Bayesian confidence score estimate $q_{i,k,c}$, capturing stable historical competence.

By setting $\lambda_f \gg \lambda_g$, the short-term average reacts rapidly while the long-term average evolves conservatively, realizing a dual-timescale adaptation mechanism. The final trust score is computed as:

$$s_{i,k,c}^{(t+1)} = \mu\, f_{i,k,c}^{(t+1)} + (1-\mu)\, g_{i,k,c}^{(t+1)} + \gamma\, \tilde{R}_i^{(t)}. \tag{14}$$

Here, μ balances short-term responsiveness against long-term stability, while the direct reward injection term $\gamma\, \tilde{R}_i^{(t)}$ ensures that strong immediate feedback can transiently influence routing before being absorbed into the moving averages. Updated scores are fed back into Stage-1 routing, closing the adaptation loop.

Intuitively, spatial errors are role-dependent: IVR failures can be compensated by 3D or scene-graph evidence, and vice versa. SpatiO operationalizes this by continuously re-allocating (i, k, c) confidence, realizing a per-category mixture-of-experts whose short-term EMA reacts quickly to distribution shifts while the Bayesian posterior resists noisy over-reaction.

5 Experiments

In this section, we present a comprehensive evaluation of SpatiO, focusing on its effectiveness for adaptive agent–role orchestration in spatial reasoning. We describe the experimental setup (Section 5.1), baselines (Section5.2), generalization (Section5.3), and inference latency (Section5.4).

5.1 Experimental Setup

Evaluation Data. We evaluate on four spatial reasoning benchmarks: STVQA-7k [4], CV-Bench [29], 3DSRBench [19], and Omni3D-Bench [6], performing inference on the full dataset for all four benchmarks.

Baselines. We compare against closed-source baselines GPT-5.2 [23] and Claude-Opus 4.6 [2], and open-source baselines Qwen3-VL-4B [3], LLaVA-4D [37], SpatialReasoner [20], SpatialRGPT [9], and Sa2VA [35]. All open-source baselines are evaluated both vanilla and LoRA-finetuned (rank r=64, α=128) on the same 150 TTO samples, controlling for data budget and isolating orchestration-level from parameter-level adaptation.

Ours (SpatiO). TTO adaptively estimates per-agent reliability across spatial categories using a stratified optimization set of 150 samples (30 per category, strictly separated from the evaluation set), yielding a fixed agent–role configuration for inference, with key hyperparameters k=0.5, μ=0.3, γ=0.3, λ_f=0.3,

Table 1: Full-dataset evaluation on 3DSRBench. Best results are **bold**, second best are underlined. **Loc.**=Location, **Ori.**=Orientation, **M.O.**=Multi-object.

Methods	3DSRBench												Overall
	Loc. above	Height higher	Loc. closer	M.O. closer	Ori. left	M.O. facing	Same dir.	Ori. front	M.O. VP	Ori. VP	Loc. next to	M.O. parallel	
Without LoRA training													
Llava-4D	51.9	48.7	58.1	58.9	50.1	44.2	50.6	56.1	20.4	24.8	57.5	56.3	49.1
Qwen-3.0-VL-4B	67.2	57.8	81.7	65.1	49.3	62.7	51.2	55.8	28.6	39.7	72.6	50.7	59.2
Sa2VA	58.2	48.6	52.2	59.4	50.1	58.7	51.2	54.1	27.1	28.6	57.5	51.9	50.5
SpatialRGPT	37.4	42.3	49.5	34.5	44.0	56.6	44.4	55.0	24.5	27.1	51.2	45.6	42.7
SpatialReasoner	69.4	51.3	73.0	57.1	49.3	50.3	47.4	51.2	29.4	31.8	69.4	48.3	52.4
With LoRA training													
Llava-4D	59.0	57.5	57.7	50.9	49.9	35.6	50.3	52.3	26.0	27.1	54.3	51.3	49.7
Qwen-3.0-VL-4B	62.4	59.1	79.9	65.7	50.1	62.4	54.4	59.3	28.3	38.2	78.2	46.9	59.1
Sa2VA	59.4	43.2	46.5	48.0	50.1	62.4	48.0	48.8	23.6	26.8	59.3	61.9	48.5
SpatialRGPT	29.0	37.9	54.3	24.5	46.9	48.5	50.3	48.5	17.5	30.2	44.3	45.2	39.8
SpatialReasoner	69.4	54.3	71.5	50.3	53.3	51.7	45.0	49.4	30.2	33.6	68.5	54.0	54.3
Gpt-5.2	68.6	**68.7**	81.3	74.3	51.0	74.3	63.4	**82.6**	38.5	**49.3**	78.2	59.6	67.2
Claude opus 4.6	73.8	63.1	77.4	78.3	53.0	**72.3**	55.4	78.3	**42.2**	48.2	60.7	59.0	63.5
SpatialMAS	73.8	64.2	84.9	74.9	58.7	48.6	62.5	74.3	35.9	43.5	79.6	75.3	67.1
SpatiO (Ours)	**75.2**	65.8	**87.3**	**76.5**	**59.4**	49.9	**63.7**	76.1	36.8	44.6	**81.5**	**77.2**	**72.4**

λ_g=0.1, T=5, β=5, temperature τ=0.7, and nucleus sampling p=0.9 for all agents (sample size ablated in Section C.1). We refer to this labeled calibration phase as *Calibrated Trust Estimation*, distinguishing it from the per-query *Test-Time Orchestration* (TTO) at inference, which requires no labels or parameter updates (full distinction in Appendix A.5).

5.2 Main Results

We report quantitative results on STVQA-7k and CV-Bench in Tab. 2, and on 3DSRBench in Tab. 1. Across all benchmarks, SpatiO consistently achieves the best overall performance, outperforming both closed-source and open-source models including LoRA-finetuned variants. On STVQA-7k, SpatiO reaches 93.1 in *size*, surpassing the previous best by 5.6 points. On CV-Bench, it achieves 79.4 in *count*, improving over the strongest baseline by 14 points. On 3DSRBench, SpatiO obtains the best performance in 8 out of 12 subcategories.

Zero-shot Performance on Omni3D-Bench. To assess out-of-distribution generalization, we evaluate on Omni3D-Bench [6] following [22]. SpatiO is optimized on 150 mixed samples (50 each from 3DSRBench, STVQA-7k, and CV-Bench), with no Omni3D-Bench samples included. As shown in Tab. 3, SpatiO achieves large gains across all output types, especially `float` (+17.1) and `int` (+20.6) over the strongest baseline, demonstrating effective heterogeneous orchestration for unseen tasks and answer formats.

5.3 Generalization to Unseen Benchmarks

We further evaluate SpatiO on MMSI-Bench [33] and MindCube [34], two benchmarks not included during TTO optimization and drawn from visual domains categorically distinct from the calibration pool (*e.g.*, autonomous driving, indoor 3D scans, egocentric video), testing whether $\mathcal{C}$ transfers beyond the calibration distribution and addressing possible benchmark-leakage concerns. SpatiO

Table 2: Comparison of methods on STVQA-7k (trained with 300 samples) and CV-Bench. Best results are **bold**, second best are underlined.

	Benchmark														
	STVQA-7k										CV-Bench				
Methods	rel.	reach	size	ori.	loc.	depth	dist.	count	exist.	Overall	count	rel.	depth	dist.	Overall
Without LoRA training															
Llava-4D	57.8	54.0	58.3	19.2	58.7	40.0	43.3	34.4	45.2	52.6	40.8	61.3	51.3	54.9	51.3
Qwen-3.0-VL-4B	83.6	88.0	70.8	73.1	**91.3**	**86.0**	76.7	56.2	**96.8**	82.4	65.0	87.7	93.8	83.5	81.3
Sa2VA	51.1	74.1	56.2	23.0	10.8	62.0	53.3	21.8	74.1	50.0	59.0	78.0	80.6	66.1	70.2
SpatialRGPT	62.8	82.0	75.0	46.2	73.9	78.0	56.7	59.4	90.3	67.1	51.0	84.4	55.3	56.5	61.4
SpatialReasoner	70.4	66.0	73.0	61.0	84.9	60.3	**86.7**	59.1	75.4	70.6	58.9	85.6	83.4	82.3	77.6
With LoRA training															
Llava-4D	63.7	65.6	54.2	22.9	60.4	52.3	46.7	30.2	39.1	57.2	61.1	61.6	82.3	70.8	68.3
Qwen-3.0-VL-4B	81.2	80.0	81.2	50.0	86.9	74.0	60.0	59.3	80.6	77.9	65.9	88.7	94.0	89.1	84.4
Sa2VA	62.3	76.0	72.9	53.8	67.4	74.0	50.0	68.8	77.4	65.3	59.0	78.0	80.7	66.2	70.2
SpatialRGPT	62.1	84.0	74.0	37.8	70.9	83.4	59.3	66.4	92.0	67.1	48.3	81.0	59.6	57.4	61.0
SpatialReasoner	60.2	78.0	87.5	38.5	63.0	66.0	53.3	62.5	71.0	63.4	61.5	86.9	82.1	83.1	77.4
Gpt-5.2	79.7	**92.0**	85.4	69.2	82.6	82.0	76.6	75.0	96.7	81.4	65.4	95.7	90.4	87.2	83.5
Claude opus 4.6	88.1	82.0	87.5	69.2	80.4	80.0	70.0	71.9	**96.8**	84.7	59.2	**94.8**	88.8	86.8	82.4
SpatiO (Ours)	**88.5**	78.0	**93.1**	**77.5**	89.0	85.4	79.4	**78.0**	92.5	**88.2**	**69.2**	91.8	**95.7**	**91.2**	**86.9**

Table 3: Comparison on Omni3D-Bench (zero-shot). Best: **bold**, second: underline.

	Omni3D-Bench		
Methods	float	int	str
Llava-4D	17.2	3.9	1.2
Qwen-3.0-VL-4B	23.2	12.3	18.0
Sa2VA	13.6	2.9	0.0
SpatialRGPT	3.1	5.4	19.2
SpatialReasoner	18.5	4.3	1.2
SpatiO	**40.3**	**32.9**	**20.0**

achieves the highest overall accuracy on both: +19.5%p on MMSI-Bench (largest gains on *Positional Relationship*, +17.8%, and *Multi-Step Reasoning*, +29.6%) and +1.67%p on MindCube. Full results, category-level analysis, and the leakage discussion are in Appendix C.2.

5.4 Inference Latency

The heterogeneous multi-agent architecture of SpatiO introduces additional inference cost relative to single-model baselines. Tab. 4 directly compares SpatiO against each single-model baseline on accuracy, average per-query latency, throughput, and peak VRAM on 3DSRBench (single A100 80GB GPU).
SpatiO improves accuracy over the strongest single-model baseline, SpatialReasoner, from 65.6% to 84.7% (+19.1%p) while *reducing* average latency from 25.8 s to 19.8 s. This is possible because the three specialist roles execute *in parallel* and

Table 4: Per-module inference latency (seconds per query, averaged over the full dataset from STVQA-7k). Specialists run in parallel; end-to-end latency is the maximum across the three selected specialists plus Head and Final Reasoning Agent.

Module	Spatial Relation	Distance Depth	Size	Counting	Orientation	Avg.
Head Agent (Qwen3-VL-4B)	2.4	2.6	2.3	2.5	2.7	2.5
Role 1 · Implicit Visual	14.2	15.8	12.1	17.3	20.6	16.0
Role 2 · Explicit 3D	2.8	8.4	2.1	2.3	7.9	4.5
Role 3 · Scene Graph	9.3	5.2	4.8	8.6	4.1	6.4
Final Reasoning Agent (DSR-7B)	4.8	5.1	3.9	4.2	4.5	4.5
End-to-end (SpatiO)	31.5	31.9	18.3	26.0	32.7	28.1

the Head Agent dynamically selects the most reliable specialist combination per query category, invoking heavyweight specialists like SpatialReasoner only when the task demands it, rather than for every query; throughput and VRAM are correspondingly higher than any individual specialist alone, reflecting the cost of running multiple specialists in parallel. Per-module and per-category latency breakdowns are provided in Appendix C.3.

6 Analysis

We further analyze TTO along three axes: (1) the contribution of each optimization component and the effect of calibration sample size, (2) how confidence scores evolve over time, and (3) whether trust scores transfer across structurally different benchmarks. Due to space constraints, full results and discussion are provided in Appendix C.1; we summarize the key findings below.

TTO component ablation. Tabs. 5 and 6 present cumulative and leave-one-out ablations of TTO on CV-Bench. Progressively adding (1) reward computation, (2) reward scaling, (3) Bayesian trust update, and (4) dual EMA each consistently improves accuracy, with the Bayesian update yielding the largest single gain by grounding agent selection in accumulated evidence. Removing the δ_i underperformance penalty (e) causes the largest single drop ($87.4 \rightarrow 72.0\%$), confirming that penalizing specialists who underperform relative to a failed consensus is critical for sharp trust differentiation. Replacing the Final Reasoning Agent with simple majority voting (d) also substantially degrades performance (77.2%), showing that conditional re-generation over the evidence pool, rather than vote counting, is responsible for much of the gain. Uniform weights (a), equivalent to disabling TTO entirely, perform only marginally better than the homogeneous baseline in Fig. 2(b), underscoring that reliability-aware orchestration is the primary source of improvement. Varying the TTO calibration sample size shows accuracy peaks at 150 samples (91.30%) before slightly degrading, motivating our default choice of 150 (30 per category; Appendix Fig. A1a). We also verify robustness to the Top-k, β, and role-assignment hyperparameters, and to substituting the agent pool or role prompts (Appendix Tabs. A3, A4).

Table 5: Ablation on each step used in TTO optimization. Results on 500 sampled CV-Bench instances with 150 optimization samples.

CV-Bench	
Methods	**Accuracy**
SpatiO	86.10
SpatiO + (step1)	73.02
SpatiO + (step2)	79.66
SpatiO + (step3)	88.12
SpatiO + (step4)	**89.86**

Table 6: Component-wise ablation on CV-Bench. Each row removes/replaces a single TTO component from the full model, isolating its individual contribution.

Variant	**Acc. (%)**
SpatiO (full)	**87.4**
(a) Uniform weights (no TTO)	75.9
(b) w/o short-term EMA	83.2
(c) w/o long-term EMA	81.5
(d) Majority voting (no Reasoner)	77.2
(e) w/o δ_i penalty	72.0

Confidence score evolution. Tracking per-agent confidence trajectories within a fixed category (Appendix Fig. A1b) shows that TTO progressively differentiates reliable agents from weaker ones within each role–category pair, realizing a per-category mixture-of-experts at test time without any parameter update.

Cross-benchmark generalization. Calibrating on 3DSRBench (3D-dominant, fine-grained categories) and evaluating zero-shot on STVQA-7k outperforms the reverse direction, calibrating on CV-Bench (2D-dominant), by +7.1% (Appendix Tab. A1, A2). This asymmetry suggests that trust profiles calibrated on harder, finer-grained 3D categories transfer more readily to simpler 2D-dominant tasks than the reverse.

7 Conclusion

In this work, we introduced SpatiO, a heterogeneous multi-agent framework for spatial reasoning that coordinates vision-language specialists with complementary inductive biases. To enable effective collaboration among agents, we proposed Test-Time Orchestration (TTO), a lightweight optimization mechanism that dynamically calibrates agent reliability during inference based on accumulated interaction evidence. Extensive experiments across diverse spatial reasoning benchmarks demonstrate that SpatiO consistently improves performance over both closed-source and open-source baselines, suggesting that dynamically orchestrating heterogeneous reasoning agents at test time is a promising direction for spatial reasoning in vision-language models without additional training.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) (RS-2026-25488668, 20%), the Institute of Information & Communications Technology Planning & Evaluation (IITP) under the Leading Generative AI Human Resources Development grant (IITP-2026-RS-2024-00397085, 20%), the Artificial Intelligence Star Fellowship Support Program to Nurture the Best Talents grant (IITP-2026-RS-2025-02304828, 40%), and the IITP-ICT Creative Consilience Program grant (IITP-2026-RS-2020-II201819, 20%), funded by the Korea government (MSIT).

References

1. Algazinov, A., Laing, M., Laban, P.: Mate: Llm-powered multi-agent translation environment for accessibility applications. arXiv preprint arXiv:2506.19502 (2025)
2. Anthropic: Claude opus 4.6. `https://www.anthropic.com/claude` (2025), accessed: 2026-03-05
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Batra, H., Tu, H., Chen, H., Lin, Y., Xie, C., Clark, R.: Spatialthinker: Reinforcing 3d reasoning in multimodal llms via spatial rewards. arXiv:2511.07403 (2025)
5. Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V.: Depth pro: Sharp monocular metric depth in less than a second (2024)
6. Brazil, G., Kumar, A., Straub, J., Ravi, N., Johnson, J., Gkioxari, G.: Omni3d: A large benchmark and model for 3d object detection in the wild. In: CVPR (2023)
7. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In: CVPR (2024)
8. Chen, Z., Lu, X., Zheng, Z., Li, P., He, L., Zhou, Y., Shao, J., Zhuang, B., Sheng, L.: Geometrically-constrained agent for spatial reasoning. ArXiv (2025)
9. Cheng, A.C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., Liu, S.: Spatialrgpt: Grounded spatial reasoning in vision-language models (2024)
10. Choi, H.K., et al.: Debate or vote: Which yields better decisions in multi-agent large language models? In: Advances in Neural Information Processing Systems (NeurIPS) (2025), spotlight
11. Choi, M., Choi, S., Kwon, M., Joung, W., Kim, J., Lee, J.: The truth stays in the family: Enhancing contextual grounding via inherited truthful heads in model lineages (2026)
12. Fung, H.L., Darvariu, V.A., Hailes, S., Musolesi, M.: Trust-based consensus in multi-agent reinforcement learning systems. arXiv preprint arXiv:2205.12880 (2022)
13. Gupta, T., Kembhavi, A.: Visual programming: Compositional visual reasoning without training. In: CVPR (2023)
14. Gupta, T., et al.: Conceptgraphs: Open-vocabulary 3d scene graphs for situated reasoning. In: CVPR (2023)
15. Hallyburton, R.S., Pajic, M.: Bayesian methods for trust in collaborative multi-agent autonomy (2024)
16. Lee, J., Chun, S., Yun, S.: Toward interactive regional understanding in vision-large language models. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 6416–6429 (2024)
17. Li, H., Li, D., Wang, Z., Yan, Y., Wu, H., Zhang, W., Shen, Y., Lu, W., Xiao, J., Zhuang, Y.: Spatialladder: Progressive training for spatial reasoning in vision-language models. ArXiv (2025)
18. Liang, J., He, R., Tan, T.: A comprehensive survey on test-time adaptation under distribution shifts. International Journal of Computer Vision **133**(1), 31–64 (2025)
19. Ma, W., Chen, H., Zhang, G., Chou, Y.C., Chen, J., de Melo, C., Yuille, A.: 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In: ICCV (2025)
20. Ma, W., Chou, Y.C., Liu, Q., Wang, X., de Melo, C., Xie, J., Yuille, A.: Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. arXiv preprint arXiv:2504.20024 (2025)

21. Ma, W., Ye, L., de Melo, C.M., Yuille, A., Chen, J.: Spatialllm: A compound 3d-informed design towards spatially-intelligent large multimodal models. In: CVPR (2025)
22. Marsili, D., Agrawal, R., Yue, Y., Gkioxari, G.: Visual agentic ai for spatial reasoning with a dynamic api. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19446–19455 (2025)
23. OpenAI: Introducing gpt-5.2. https://openai.com/index/introducing-gpt-5-2/ (2025), accessed: 2026-03-05
24. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision (2023)
25. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv:2408.00714 (2024)
26. Stogiannidis, I., McDonagh, S., Tsaftaris, S.A.: Mind the gap: Diagnosing spatial reasoning failures in vision-language models. ArXiv (2025)
27. Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., Hardt, M.: Test-time training with self-supervision for generalization under distribution shifts. In: International conference on machine learning. pp. 9229–9248. PMLR (2020)
28. Surís, D., Menon, S., Vondrick, C.: ViperGPT: Visual inference via Python execution for reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11854–11864 (2023)
29. Tong, P., Brown, E., Wu, P., Woo, S., Iyer, A.J.V., Akula, S.C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms (2024)
30. Wang, J., Wang, J., Athiwaratkun, B., Zhang, C., Zou, J.: Mixture-of-agents enhances large language model capabilities. arXiv preprint arXiv:2406.04692 (2024)
31. Xue, Q., Liu, W., Wang, S., Wang, H., Wu, Y., Gao, W.: Reasoning path and latent state analysis for multi-view visual spatial reasoning: A cognitive science perspective. ArXiv (2025)
32. Yang, S., Li, Y., Lam, W., Cheng, Y.: Multi-llm collaborative search for complex problem solving. arXiv preprint arXiv:2502.18873 (2025)
33. Yang, S., Xu, R., Xie, Y., Yang, S., Li, M., Lin, J., Zhu, C., Chen, X., Duan, H., Yue, X., et al.: Mmsi-bench: A benchmark for multi-image spatial intelligence (2025)
34. Yin, B., Wang, Q., Zhang, P., Zhang, J., Wang, K., Wang, Z., Zhang, J., Chandrasegaran, K., Liu, H., Krishna, R., et al.: Spatial mental modeling from limited views. In: Structural Priors for Vision Workshop at ICCV'25 (2025)
35. Yuan, H., Li, X., Zhang, T., Sun, Y., Huang, Z., Xu, S., Ji, S., Tong, Y., Qi, L., Feng, J., et al.: Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. arXiv arXiv:2501.04001 (2025)
36. Zhang, W., Zhou, Z., Zeng, X., Liu, X., Fang, J., Gao, C., Li, Y., Cui, J., Chen, X., Zhang, X.P.: Open3d-vqa: A benchmark for comprehensive spatial reasoning with multimodal large language model in open space (2025)
37. Zhou, H., Lee, G.H.: Llava-4d: Embedding spatiotemporal prompt into lmms for 4d scene understanding. arXiv:2505.12253 (2025)