SCOUTCREW: COORDINATING MULTI-VIEW EXPLORATION AND EVIDENCE SYNTHESIS FOR SPATIAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial reasoning often requires information that is not fully visible in the supplied images. Additional viewpoints can reveal useful cues, but their value depends on what they show and how their findings are combined. We present SCOUTCREW, a coordinated multi-agent framework that plans observations together with their roles in the inference. A Coordinator assigns observation subtasks and constructs an Investigator Relation Graph (IRG). Complement relations connect tasks that supply different premises, while CrossCheck relations connect tasks that examine the same spatial proposition. Investigators execute their plans through generative view synthesis. They then exchange reports and observation context along the IRG to resolve missing information and revise their findings. An Analyst assembles the revised reports into a chain of evidence (CoE), connecting observations and intermediate conclusions to the final answer. ScoutCrew improves overall accuracy across all evaluated backbones on SAT-Real and multi-image MMSI-Bench. Gains over direct inference reach up to 14.0 and 9.5 percentage points, respectively. These results support coordinating observation acquisition and evidence composition for spatial reasoning from limited views.

1 INTRODUCTION

Spatial reasoning is the ability to understand and infer the positions, orientations, and relationships of objects in space (Chen et al., 2024; Ray et al., 2025). It supports robotic navigation and object manipulation (Huang et al., 2025). In visual spatial reasoning, the available views may not contain enough information to answer a question (Yang et al., 2026; 2025b). Consider the question in Figure 1: *which mug is to the right of the laptop and closer to it than the red notebook?* Here, left and right are defined relative to the input view. This view may reveal the directional relation while leaving the distance comparison ambiguous. Additional viewpoints can help resolve that comparison, but answering the question requires connecting findings about the same objects across views.

Recent work improves spatial reasoning through richer training data, explicit spatial representations, and synthesized observations. Spatial supervision and synthetic training data strengthen geometric understanding and reasoning about spatial changes (Chen et al., 2024; Cheng et al., 2024; Ray et al., 2025). Cognitive maps and 3D sandboxes provide structured representations of scene layout and object geometry to support spatial reasoning (Yang et al., 2025a; Liu et al., 2026). Tool-using agents also obtain geometric information through 3D reconstruction and specialized perception tools (Luo et al., 2026; Chen et al., 2026). View synthesis provides images from additional camera positions (Zhou et al., 2025). World-model exploration uses this capability to search camera trajectories and adapt visual exploration to the question (Yang et al., 2025b; Yu et al., 2026). ViSA evaluates imagined views using frame-anchored spatial assertions (Jha et al., 2025). Together, these approaches improve what models know about space, how they represent it, and what they can inspect.

Beyond observation acquisition, self-consistency aggregates answers across sampled reasoning paths (Wang et al., 2023), and multi-agent debate revises answers through interaction (Du et al., 2024). GPTSwarm represents collaborating agents as computational graphs (Zhuge et al., 2024), while G-Designer learns task-adaptive communication topologies (Zhang et al., 2025). These approaches

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

motivate organizing how reasoning outputs are combined. A detailed discussion of related work is provided in Appendix A.

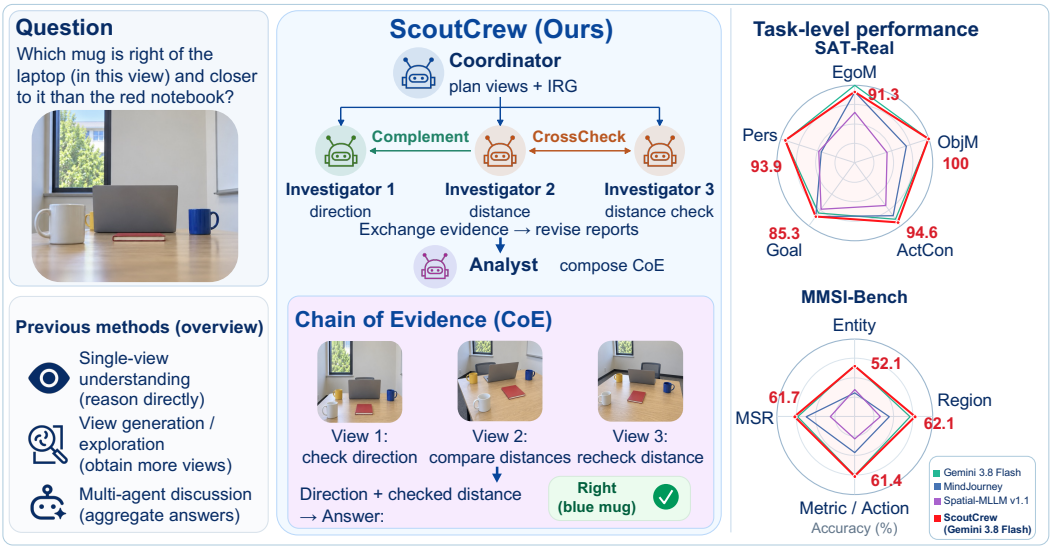

Figure 1: **Comparing our method with existing approaches.** The mug question combines a directional relation with a distance comparison. Left and right refer to the input view, the answer *Right* identifies the blue mug shown on the right. The right panel compares ScoutCrew with existing methods, with ScoutCrew achieving the highest overall accuracy on both benchmarks.

For a particular question, however, useful reasoning depends on both *where to look* and *how to use the resulting observations*. Repeated views can preserve the same occlusion, and unnecessary visual imagination can introduce misleading evidence (Yu et al., 2026). In this example, one view may establish that a candidate mug is to the right of the laptop in the input-view reference frame. Another may establish that the same mug is closer to the laptop than the red notebook is. Together, these findings establish that the mug satisfies both conditions. A further view can check either relation, but rechecking one does not establish the other. The value of an observation therefore depends on whether it supplies missing information or checks information already obtained.

We introduce SCOUTCREW, a coordinated multi-agent framework that plans observations together with their roles in the inference (Figure 2). The Coordinator Agent decomposes the question into observation subtasks and jointly produces scouting plans and an *Investigator Relation Graph* (IRG). Each node represents an Investigator Agent. Complement edges connect tasks that supply different premises of a joint inference, while CrossCheck edges connect tasks that examine the same spatial proposition. Each Investigator follows its planned trajectory through generative view synthesis. It returns an investigation report containing a tentative local answer, spatial findings, and unresolved evidence gaps. This joint planning gives each investigation a specific purpose and defines how its findings will contribute to crew reasoning.

The IRG guides information exchange and report revision before the final answer is derived. Investigators share reports together with camera trajectories and selected views, providing the context needed to interpret each finding. They use complementary information to fill missing premises and cross-checks to support or correct their conclusions. The Scout Analyst Agent then combines the revised reports into a *chain of evidence* (CoE). Following the IRG, it identifies findings that can be combined, examines the outcomes of cross-checks, and determines whether the required evidence gaps have been resolved. The resulting chain connects observations through intermediate conclusions to the final answer.

Experiments on SAT-Real and multi-image MMSI-Bench show overall accuracy improvements across all evaluated backbones. The mean paired gains are 7.0 percentage points across seven SAT-Real backbones and 5.2 points across six MMSI-Bench backbones. With Gemini 3.8 Flash, SCOUTCREW reaches 92.7% and 59.1%, respectively, the highest overall accuracies among the evaluated configurations. **Our contributions are threefold:**

- We propose a **coordinated multi-agent framework** for spatial reasoning that jointly plans observation subtasks and their relationships, giving each investigation a defined role in answering the question.
- We develop **IRG-guided information exchange and report revision** to combine complementary findings and cross-check spatial claims. CoE synthesis organizes the revised findings into an explicit answer derivation.
- We demonstrate **consistent overall accuracy gains** across closed- and open-weight backbones on SAT-Real and MMSI-Bench.

2 MOTIVATION

Additional viewpoints help spatial reasoning only when they provide information needed for the answer. In the mug question in Figure 1, a candidate must be to the right of the laptop in the input-view reference frame and closer to it than the red notebook is. Repeatedly observing the directional relation cannot resolve an uncertain distance comparison. Observation planning should therefore identify the missing premises and select viewpoints that can reveal them.

Observations serve two distinct roles: *supplying complementary premises* and *checking the same proposition*. The directional and distance findings jointly establish whether the mug satisfies both conditions. Another view of the distance relation can instead support or challenge that finding. Combining these observations requires matching object identities and accounting for their reference frames. Our ablations support the importance of this coordination: fixed observation tasks achieve 60.7% accuracy on SAT-Real with GPT-4o, below direct inference at 64.7% (Table 3). Additional observations alone do not ensure better answers.

We therefore plan each observation together with its role in the inference. The *Investigator Relation Graph* (IRG) connects observation subtasks through Complement and CrossCheck relations. These relations guide information exchange so that Investigators can fill missing premises and reassess their findings. The Analyst organizes the revised findings into a *chain of evidence* (CoE), connecting observations and intermediate conclusions to the final answer.

3 METHOD

3.1 OVERALL ARCHITECTURE

Given a spatial question q and a limited set of N input images $\mathbf{X} = \{X_n\}_{n=1}^{N}$, SCOUTCREW produces an answer y by acquiring and combining observation evidence from different viewpoints. The framework jointly plans what to observe and how the resulting findings contribute to the answer. Figure 2 presents the overall pipeline. SCOUTCREW consists of three agent roles: a Coordinator $\mathbf{CoordA}$, K Investigators $\{\mathbf{InvA}_i\}_{i=1}^{K}$, and an Analyst $\mathbf{AnalystA}$. The Coordinator assigns observation plans and constructs an *Investigator Relation Graph* (IRG) $\mathcal{G}$. Its edges specify whether two tasks provide complementary premises or examine the same spatial proposition. Each Investigator executes its plan, and returns a planned view X_i^{plan} and an investigation report r_i (Section 3.2). Investigators then exchange information along the IRG and revise their reports. The Analyst combines the revised findings into a *chain of evidence* (CoE) $\mathcal{C}$, which connects observations and intermediate conclusions to the final answer (Section 3.3).

3.2 SPATIAL SCOUTING CREW

3.2.1 COORDINATOR

The Coordinator decomposes the question into observation subtasks whose findings can be combined or cross-checked. Given q and $\mathbf{X}$, it jointly produces K plans and the IRG:

$$(\{p_i\}_{i=1}^{K}, \mathcal{G}) = \mathbf{CoordA}(q, \mathbf{X}), \quad p_i = (X_i^{\text{init}}, d_i, \tau_i). \tag{1}$$

Here $X_i^{\text{init}} \in \mathbf{X}$ is the initial image, d_i describes the desired view, and τ_i specifies the camera actions needed to reach it. The description identifies the objects and spatial information to inspect without prescribing the answer. For example, it may request a view showing the candidate mug,

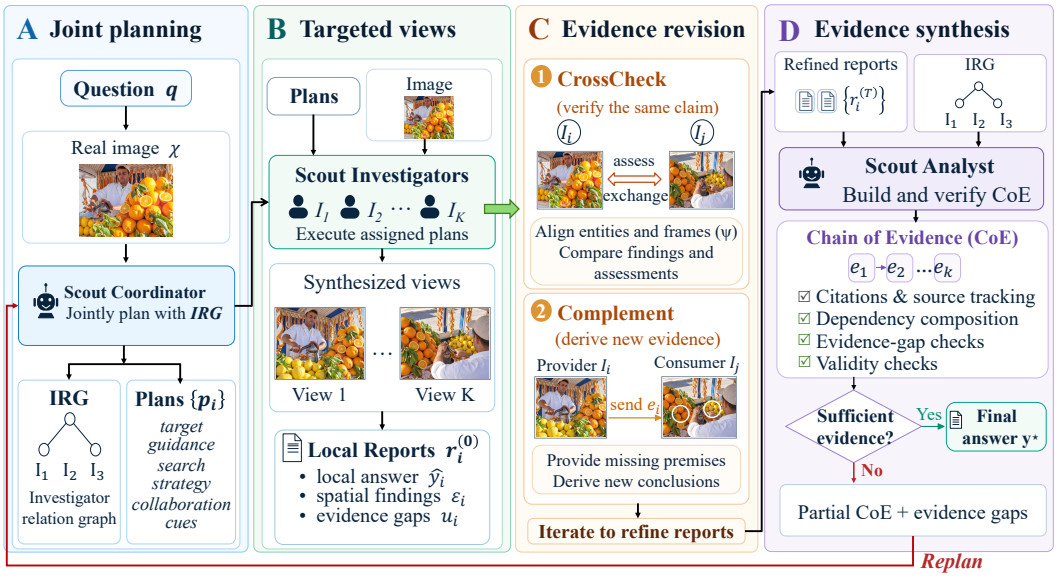

Figure 2: **ScoutCrew overview.** The Coordinator plans observations and their evidence relations together. Investigators exchange premises or independent assessments according to the IRG and use them to revise their reports. The Analyst orders the resulting dependencies into a CoE, while unresolved evidence needs return to planning.

laptop, and red notebook to compare the mug's and notebook's distances to the laptop. The trajectory $\tau_i = (a_{i,1}, \ldots, a_{i,T_i})$ contains T_i actions that move the camera from the pose of the initial view to the planned viewpoint. Actions use fixed translation and rotation increments. Horizontal scouting uses forward translation and left or right yaw, auxiliary observation plans can also include pitch adjustments, as in Figure 3.

The IRG specifies how the planned subtasks contribute to a shared inference:

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \rho), \qquad \mathcal{V} = \{\mathbf{InvA}_i\}_{i=1}^{K}, \quad \rho : \mathcal{E} \to \{\text{Complement}, \text{CrossCheck}\}. \tag{2}$$

Each node represents an Investigator assigned to one plan. The edge set $\mathcal{E}$ connects related Investigators, and ρ assigns a relation label to each edge. Complement connects tasks that supply different premises of a joint inference. CrossCheck connects tasks that examine the same spatial proposition from different viewpoints. These relations guide subsequent information exchange and report revision. For the mug question in Figure 1, Investigator $\mathbf{InvA}_1$ checks whether a candidate mug is to the right of the laptop in the input-view reference frame. Investigator $\mathbf{InvA}_2$ checks whether the same mug is closer to the laptop than the red notebook is. A Complement edge connects these tasks because their findings jointly determine whether the mug satisfies both conditions. Investigator $\mathbf{InvA}_3$ checks the same distance comparison from another viewpoint, forming a CrossCheck edge with $\mathbf{InvA}_2$. Its observation may support or challenge the distance finding. The relation remains CrossCheck because the label specifies what to verify, while the reports record the outcome.

3.2.2 Investigator

Each Investigator $\mathbf{InvA}_i$ executes its assigned plan and independently examines the scene to answer the question. Since static-image inputs do not permit physical camera movement, we use Stable Virtual Camera (SVC) (Zhou et al., 2025) to generate observations. We denote its rendering interface by $\mathcal{W}_{\text{SVC}}(X^{init}, \tau)$, which converts an action sequence τ into camera poses relative to a reference image X^{init} and returns the terminal generated view. Camera-action settings and SVC rendering are described in Appendices D and E, respectively. The Investigator executes the planned trajectory from the initial image:

$$X_i^{\text{plan}} = \mathcal{W}_{\text{SVC}}\left(X_i^{\text{init}}, \tau_i\right). \tag{3}$$

The Investigator examines q, the input images $\mathbf{X}$, and the planned observation X_i^{plan} to produce its report:

$$r_i = \mathbf{InvA}_i\left(q, \mathbf{X}, X_i^{plan}\right) = (\hat{y}_i, \mathcal{E}_i, \mathcal{U}_i^{\text{gap}}). \tag{4}$$

The report contains a *tentative local answer* $\hat{y}_i$, a set of *spatial findings* $\mathcal{E}_i$, and a set of *evidence gaps* $\mathcal{U}_i^{\text{gap}}$. A finding states an observed or inferred spatial proposition and identifies the relevant objects and reference frame. Observed findings refer to their supporting view, while derived findings also cite their premises. An evidence gap identifies an unresolved proposition and the observation needed to resolve it. In the mug example, an Investigator may establish that a candidate mug is to the right of the laptop but lack evidence for the distance comparison. Its report records the directional relation in the input-view reference frame as a finding, and the unresolved distance comparison as an evidence gap. Each Investigator produces its initial report independently before information exchange.

3.3 Crew Reasoning with Chain of Evidence

Crew reasoning uses the IRG to connect findings across reports. Investigators first exchange information according to their task relations. They then revise their reports, and the Analyst assembles the supported findings into a *chain of evidence* (CoE).

3.3.1 IRG-Guided Information Exchange

During information exchange, each IRG node holds an Investigator, its selected observation, and its report. The edge labels determine which Investigators communicate and how they use the received information. To make the observation context explicit, each message includes the report and its acquisition context:

$$m_i = \left(X_i^{\text{init}}, \tau_i, X_i^{plan}, r_i\right). \tag{5}$$

The initial image and planned trajectory describe how the terminal generated view was obtained. This context helps the recipient interpret the reported spatial relations from the appropriate viewpoint.

Complement supplies a finding needed by another Investigator. In the mug example, a report establishing the distance comparison can be sent to an Investigator that has established the directional relation. The recipient matches the mug and laptop across reports and interprets the directional finding in the input-view reference frame. It then combines both findings to determine whether that mug satisfies the question's two conditions.

CrossCheck exchanges reports between Investigators examining the same proposition. For the mug example, $\mathbf{InvA}_2$ and $\mathbf{InvA}_3$ reassess whether the same mug is closer to the laptop than the red notebook is. Each uses the other's observation, trajectory, and report to check object identities and distinguish scene distances from apparent distances in the image. The received evidence may support the current finding or warrant a correction. The two forms of information exchange are expressed as

$$\text{Complement} : \mathbf{InvA}_i \xrightarrow{m_i} \mathbf{InvA}_j, \quad \text{CrossCheck} : \mathbf{InvA}_i \underset{m_j}{\overset{m_i}{\rightleftarrows}} \mathbf{InvA}_j. \tag{6}$$

For Complement, the message goes to the Investigator that needs the additional premise. CrossCheck is reciprocal, allowing both Investigators to reassess their reports.

3.3.2 Chain-of-Evidence Synthesis

Each Investigator revises its report using its own observation context m_i and the received messages $\mathcal{M}_i$:

$$r_i' = \mathbf{InvA}_i^{\text{revise}}\left(q, \mathbf{X}, m_i, \mathcal{M}_i\right) = \left(\hat{y}_i', \mathcal{E}_i', \mathcal{U}_i^{\text{gap}\prime}\right). \tag{7}$$

Complementary information can fill a missing premise, while a cross-check can support or correct a finding. The Investigator updates its local answer and removes only the gaps resolved by the new information. A finding is marked as *active* when its observations and premises support it without an unresolved conflict. It remains *open* when a required premise is missing or a disagreement remains unresolved. A finding replaced by a supported correction is marked as *withdrawn*. The Analyst receives $\mathcal{G}$ and all revised reports. It follows Complement edges to collect findings that jointly support an inference and CrossCheck edges to examine how those findings were confirmed or corrected. It also reads the updated evidence-gap fields to identify any missing premises. The Analyst then orders the supported findings into a *Chain of Evidence* (CoE), placing each intermediate conclusion after the premises used to derive it:

$$\begin{aligned} \mathcal{C} &= \mathbf{AnalystA}_{\text{synth}}\left(q, \mathbf{X}, \mathcal{G}, \{r_i'\}_{i=1}^{K}\right) = (h_1, \ldots, h_L), \\ h_k &= (\mathcal{P}_k \Rightarrow c_k), \qquad y = \mathbf{AnalystA}_{\text{answer}}\left(q, \mathbf{X}, \mathcal{C}\right). \end{aligned} \tag{8}$$

Table 1: SAT-Real results. All scores are accuracies in percent. Colored bands identify model groups, and teal rows report SCOUTCREW with the indicated backbone. Bold and underlined Overall values mark the highest and second-highest scores across all listed configurations.

Method	Spatial task accuracy (%, ↑)					Overall Acc.↑
	EgoM	ObjM	ActCon	Goal	Pers	
Baseline Models						
GPT-4o	69.6	60.9	75.7	50.0	66.7	64.7
GPT-4.1	87.0	87.0	78.4	76.5	51.5	74.7
GPT-5.6 Sol	100.0	87.0	83.8	76.5	81.8	84.7
Gemini 3.8 Flash	100.0	100.0	89.2	79.4	93.9	<u>91.3</u>
Claude Fable 5	87.0	60.9	75.7	67.6	57.6	69.3
Qwen3-VL-235B	100.0	73.9	73.0	82.4	39.4	72.0
InternVL3.5-14B	56.5	56.5	78.4	67.6	57.6	64.7
Spatially specialized models						
SpatialRGPT	52.2	87.0	64.9	67.6	48.5	63.3
Spatial-MLLM v1.1	65.2	43.5	67.6	73.5	48.5	60.7
AVIC / GPT-4.1	82.6	65.2	75.7	73.5	48.5	68.7
MindJourney / GPT-4.1	91.3	69.6	83.8	85.3	45.5	74.7
Ours: backbone + SCOUTCREW						
GPT-4o	100.0	69.6	89.2	76.5	60.6	78.7
GPT-4.1	100.0	91.3	83.8	85.3	78.8	86.7
GPT-5.6 Sol	95.7	100.0	75.7	88.2	84.8	87.3
Gemini 3.8 Flash	91.3	100.0	94.6	85.3	93.9	**92.7**
Claude Fable 5	87.0	60.9	83.8	79.4	69.7	76.7
Qwen3-VL-235B	100.0	69.6	86.5	76.5	60.6	78.0
InternVL3.5-14B	73.9	73.9	73.0	73.5	57.6	70.0

Here L is the number of inference steps. The premise set $\mathcal{P}_k$ contains findings from the revised reports or conclusions from earlier steps, and c_k is the conclusion derived at step k. Each step identifies the reports supporting its premises. Open or withdrawn findings cannot be used as established premises.

For the mug question in Figure 1, the chain combines a directional finding with a cross-checked distance comparison for the same candidate mug. The first establishes that the mug is to the right of the laptop in the input-view reference frame. The second establishes that it is closer to the laptop than the red notebook is. Together, these premises identify a mug satisfying both conditions. If either relation remains unresolved and acquisition budget remains, the Analyst returns the partial chain and missing observation requirement to the Coordinator. When the budget is exhausted, the system returns the cached input-based answer with an incomplete-evidence status (Figure 2, Appendix D). Figure 3 provides a concrete store-entrance example, showing how observation context and peer reports correct a local answer and support the final evidence chain.

4 EXPERIMENTS

4.1 EVALUATION PROTOCOL

We evaluate SCOUTCREW on **SAT-Real** and **MMSI-Bench** (Ray et al., 2025; Yang et al., 2026). SAT-Real evaluates spatial reasoning from supplied images, while MMSI-Bench focuses on relations across multiple input views. SAT-Real covers egocentric movement (EgoM), object movement (ObjM), action consequences (ActCon), goal direction (Goal), and perspective (Pers). For MMSI-Bench, we report Entity, Region, and Metric/Action accuracy as averages over three, three, and four relation categories, respectively, alongside multi-step reasoning (MSR). Overall denotes the benchmark aggregate. Accuracies are reported as percentages and gains as percentage points.

Tables 1 and 2 compare direct inference with SCOUTCREW using the same answering backbone and benchmark split, across seven backbones on SAT-Real and six on MMSI-Bench. We average the paired differences in Overall accuracy with equal weight per backbone. SpatialRGPT (Cheng et al., 2024) and Spatial-MLLM (Wu et al., 2025) serve as *spatially specialized baselines*. MindJourney (Yang et al., 2025b) and AVIC (Yu et al., 2026), both using GPT-4.1, provide comparisons with alternative *test-time viewpoint exploration methods*. These reference methods are excluded from the paired averages. The ablation and cost studies use GPT-4o on SAT-Real.

Table 2: MMSI-Bench results. All scores are percentages. Entity, Region, and Metric/Action average three, three, and four relation categories, respectively. MSR denotes multi-step reasoning, and Overall is the benchmark aggregate. Teal rows report SCOUTCREW. Bold and underlined Overall values mark the highest and second-highest scores across all listed configurations.

Method	Grouped diagnostic accuracy (%, ↑)			Multi-step	Overall Acc.↑
	Entity	Region	Metric/Action	MSR	
Baseline Models					
GPT-4o	26.2	29.6	34.1	30.8	30.3
GPT-4.1	30.8	30.5	32.8	28.8	30.9
GPT-5.6 Sol	42.8	40.5	49.3	48.5	45.2
Gemini 3.8 Flash	52.7	57.0	60.5	58.6	<u>57.2</u>
Qwen3-VL-235B	29.1	34.2	30.7	31.8	31.4
InternVL3.5-14B	24.5	28.4	27.1	31.3	27.6
Spatially specialized models					
SpatialRGPT	26.8	25.3	27.9	25.8	26.5
Spatial-MLLM v1.1	27.9	26.5	22.9	25.3	25.6
AVIC / GPT-4.1	27.8	34.8	40.0	32.3	33.8
MindJourney / GPT-4.1	24.4	35.6	36.7	50.0	34.0
SCOUTCREW with each backbone					
GPT-4o	34.7	34.0	39.3	33.0	36.0
GPT-4.1	27.8	39.5	38.7	35.6	35.4
GPT-5.6 Sol	47.9	57.6	52.5	58.1	53.7
Gemini 3.8 Flash	52.1	62.1	61.4	61.7	**59.1**
Qwen3-VL-235B	42.6	40.3	41.4	38.5	40.9
InternVL3.5-14B	24.9	31.1	27.1	31.3	28.4

4.2 MAIN RESULTS ACROSS MODEL FAMILIES

SCOUTCREW improves overall **SAT-Real** accuracy on all seven backbones, with a mean gain of **7.0 points** (Table 1). The gains span closed- and open-weight models: GPT-4o and GPT-4.1 improve by 14.0 and 12.0 points, while Qwen3-VL and InternVL3.5 improve by 6.0 and 5.3 points. GPT-4.1 improves in all five task categories, most notably on Perspective, where accuracy rises from 51.5% to 78.8%, Qwen3-VL gains 21.2 points on the same category. Crew reasoning also benefits the stronger direct baselines, adding 2.6 points to GPT-5.6 Sol and 1.4 points to Gemini 3.8 Flash. The latter reaches 92.7%, the highest overall accuracy in the comparison.

The gains extend to multi-image reasoning: all six backbones improve on **MMSI-Bench**, by **5.2 points** on average (Table 2). Qwen3-VL and GPT-5.6 Sol gain the most, with increases of 9.5 and 8.5 points. Their category results show improvements in relating spatial information: Qwen3-VL gains 13.5 points on Entity relations, while GPT-5.6 Sol gains 17.1 points on Region relations and 9.6 points on multi-step reasoning. Gemini 3.8 Flash achieves the highest overall accuracy at 59.1%. These results show that crew reasoning remains useful when multiple views are already supplied as input. With GPT-4.1 as the shared backbone, SCOUTCREW outperforms both viewpoint exploration methods in overall accuracy. It reaches 86.7% on SAT-Real, compared with 74.7% for MindJourney and 68.7% for AVIC. The corresponding MMSI-Bench accuracies are 35.4%, 34.0%, and 33.8%. The advantage is therefore larger on SAT-Real. MindJourney retains the strongest multi-step reasoning score among these three methods on MMSI-Bench, indicating that the overall ranking does not hold for every spatial task.

4.3 VISUALIZING CREW REASONING ON A SPATIAL QUESTION

Figure 3 follows SAT-Real question 123: whether turning right by 90° would face a store entrance. Three Investigators inspect a right-turn view, a left-turn view, and an auxiliary view obtained by adjusting pitch and moving forward. Their initial answers are *No*, *Yes*, and *Yes*. The right-turn view shows the street, whereas the left-turn view reveals the entrance. Because the question concerns a right turn from the initial heading, visibility in the left-turn view alone cannot justify *Yes*.

Investigator 1 receives reports from Investigators 2 and 3 and retains *No*. Investigator 3 receives reports from Investigators 1 and 2, locates the entrance to the left of the initial heading, and corrects *Yes* to *No*. Investigator 2 receives no review message and retains *Yes*. The Analyst connects the entrance location with the left- and right-turn observations in the initial reference frame to conclude

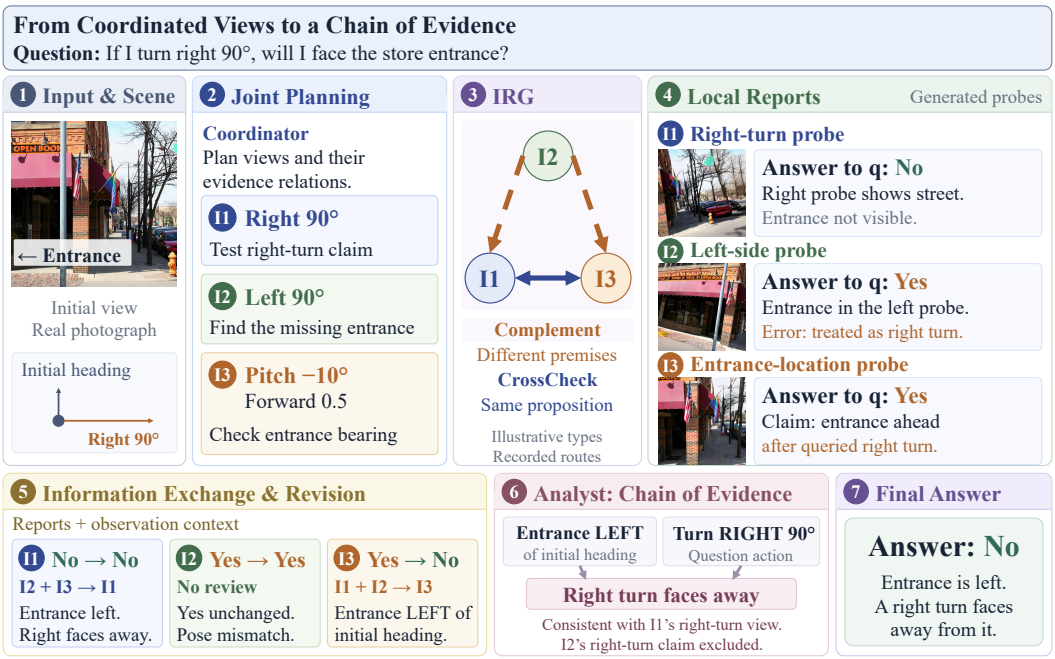

Figure 3: **From coordinated exploration to a chain of evidence.** The panels connect the input question, observation plans, Investigator relations, reports, report revision, evidence synthesis, and final answer. IRG labels describe the roles of exchanged evidence, and the CoE summarizes the recorded final explanation. Peer evidence corrects Investigator 3 from *Yes* to *No*.

No. This CoE shows how complementary observations and cross-checks resolve an incorrect local interpretation before final synthesis.

4.4 Ablations of Crew Organization and Report Integration

Table 3 tests how task planning, multiple Investigators, and evidence integration contribute to the full crew. Uncoordinated investigation assigns three fixed camera tasks without Coordinator planning or IRG checks. It reaches only 60.7% accuracy, below direct inference at 64.7%, whereas the full crew reaches 78.7%. A single Investigator reaches 68.0%. Together, these comparisons show that the benefit of multiple Investigators depends on how their tasks and findings are coordinated.

Table 3: Role and component ablation on SAT-Real with GPT-4o. Variants examine question-specific planning, multiple Investigators, cross-view verification, and CoE synthesis. Accuracy is in percent. Δ reports the difference from the full system.

Variant	Crew components				Accuracy	
	Plan	Multi	Check	CoE	Acc.↑	Δ
Direct	✗	✗	✗	✗	64.7	−14.0
Uncoordinated	✗	✓	✗	✓	60.7	−18.0
Single Investigator	✓	✗	✗	✓	68.0	−10.7
w/o IRG checks	✓	✓	✗	✓	75.3	−3.3
Voting	✓	✓	✗	✗	75.3	−3.3
Merged roles	✓	✓	✓	▲	69.3	−9.4
Full ScoutCrew	✓	✓	✓	✓	**78.7**	–

Table 4: Evidence-integration study on SAT-Real with GPT-4o. Alternative strategies use frozen initial reports, with the full-system run as the reference. Fixes, breaks, and override precision characterize the benefits and errors introduced by answer revision.

Strategy	Acc.	Revisions		Prec.
	(%)↑	Fixes↑	Breaks↓	(%)↑
Text concatenation	73.3	31	18	63.3
Majority vote	73.3	31	18	63.3
Image-aware Single Analyst	70.7	31	22	58.5
All-to-all discussion	72.0	32	21	60.4
One-way critique	71.3	33	23	58.9
w/o Source Correction	61.3	23	28	45.1
w/o Alignment Check	62.7	25	28	47.2
Full ScoutCrew	**78.7**	**35**	**14**	**71.4**

Cross-view checks and a dedicated synthesis role further improve accuracy. Disabling IRG-guided checks while retaining planning and graph construction reduces accuracy to 75.3%. Vote resolution, which selects among initial Investigator answers before exchange and CoE synthesis and uses confidence to break ties, also reaches 75.3%. Both fall 3.3 points below the full crew. Assigning the Analyst's responsibilities to the Coordinator yields 69.3%, a 9.4-point drop, supporting the separation of investigation planning from final evidence synthesis.

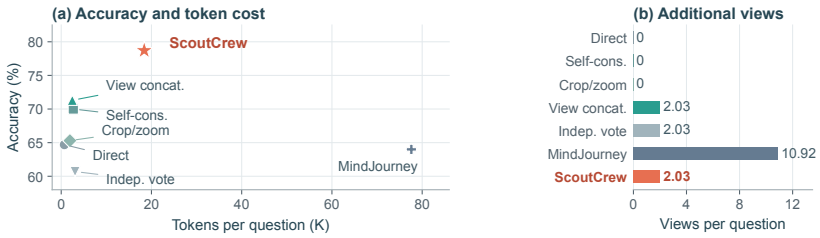

Figure 4: **Inference cost on SAT-Real with GPT-4o.** (a) Accuracy versus mean token use. (b) Mean additional selected views, excluding input images and crop/zoom operations.

Table 4 compares ways to combine Investigator reports. The alternative strategies use frozen initial reports, with the existing full-system run as the reference. Report concatenation receives text only, the Single Analyst also receives the images. All-to-all discussion shares all reports for one round, while unidirectional critique assigns each Investigator one peer without a reply.

SCOUTCREW achieves the highest accuracy and override precision, with 35 fixes and 14 breaks. Its override precision is $35/(35+14) = 71.4\%$, compared with 63.3% for majority voting. Both make 49 correctness-changing revisions, but the full crew produces four more fixes and four fewer breaks. All-to-all discussion reaches 72.0% accuracy with 21 breaks, compared with 78.7% and 14 breaks for SCOUTCREW. The advantage lies in making better-supported revisions, exchanging more reports alone does not achieve the same balance of fixes and breaks. Source-dependence correction and cross-view alignment checks limit harmful revisions. The former caps the total confidence weight of reports within a shared-anchor group at one, the latter checks that a peer's assessment refers to the same entities, relation, and reference frame. Removing either check doubles the number of breaks from 14 to 28 and lowers accuracy to 61.3% or 62.7%, respectively. In both variants, breaks outnumber fixes. The resulting loss connects these checks directly to the reliability of evidence integration: shared sources must be accounted for, and reports must refer to compatible spatial claims before they are combined.

4.5 INFERENCE COST

SCOUTCREW achieves the highest accuracy in Figure 4 while using substantially fewer resources than MindJourney. It reaches 78.7% with 18.4K tokens and 2.03 additional views per question, compared with MindJourney's 64.0%, 77.6K tokens, and 10.92 views. Thus, acquiring more views does not guarantee a better answer in this comparison. The comparison with simpler integration strategies isolates a *different trade-off*. Selected-view concatenation and independent Investigator voting use the same mean of 2.03 additional views as SCOUTCREW, but reach 71.3% and 60.7% accuracy at 2.46K and 3.07K tokens, respectively. The full crew spends more tokens on coordination, verification, and synthesis, gaining 7.4 points over concatenation. Its advantage therefore comes with additional reasoning cost even when the number of acquired views is unchanged.

5 CONCLUSION

We presented SCOUTCREW, a coordinated multi-agent framework for spatial reasoning from limited input views. The framework jointly plans observation subtasks and their relations through an Investigator Relation Graph (IRG). Investigators acquire targeted observations, and exchange information to combine complementary findings and cross-check spatial claims. The Analyst organizes the revised reports into a chain of evidence (CoE) that connects observations and intermediate conclusions to the final answer. Experiments on SAT-Real and MMSI-Bench show overall accuracy improvements across all evaluated backbones. These results support the central principle of SCOUTCREW: planning what to observe together with how the resulting information will be combined and checked.

AI USE STATEMENT

We used AI tools to assist with language editing and polishing, as well as literature retrieval and discovery of potentially relevant work. The authors retain full responsibility for the manuscript,

including the accuracy of its content, the validity of its claims, and the verification of all cited references.

Reproducibility Statement

We describe the architecture, evaluation protocol, and rendering settings throughout the main paper and appendix. Section 3 specifies the ScoutCrew architecture, including observation planning, IRG-guided evidence exchange, report refinement, and Chain-of-Evidence synthesis. Section 4 describes the evaluation benchmarks, model backbones, baselines, and evaluation protocol. Camera actions, trajectory discretization, and renderer settings are provided in Appendix D of the supplementary material. The SVC interface is detailed in Appendix E. Additional implementation details, agent prompts, model configurations, and evaluation settings are provided in the supplementary material.

References

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

Siyi Chen, Mikaela Angelina Uy, Chan Hee Song, Faisal Ladhak, Adithyavairavan Murali, Qing Qu, Stan Birchfield, Valts Blukis, and Jonathan Tremblay. SpaceTools: Tool-augmented spatial reasoning via double interactive RL. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37109–37120, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Chen_SpaceTools_Tool-Augmented_Spatial_Reasoning_via_Double_Interactive_RL_CVPR_2026_paper.html.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. SpatialRGPT: Grounded spatial reasoning in vision-language models. arXiv preprint arXiv:2406.01584, 2024. URL https://arxiv.org/abs/2406.01584.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 11733–11763. PMLR, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Multimodal spatial language maps for robot navigation and manipulation. *The International Journal of Robotics Research*, 2025. doi: 10.1177/02783649251351658. URL https://journals.sagepub.com/doi/10.1177/02783649251351658.

Saurav Jha, M. Jehanzeb Mirza, Wei Lin, Shiqi Yang, and Sarath Chandar. Probing the effectiveness of world models for spatial reasoning through test-time scaling. arXiv preprint arXiv:2512.05809, 2025. URL https://arxiv.org/abs/2512.05809.

Yifan Liu, Fangneng Zhan, Kaichen Zhou, Yilun Du, Paul Pu Liang, and Hanspeter Pfister. Abstract 3D perception for spatial intelligence in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38647–38656, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Liu_Abstract_3D_Perception_for_Spatial_Intelligence_in_Vision-Language_Models_CVPR_2026_paper.html.

Zhanpeng Luo, Ce Zhang, Silong Yong, Cunxi Dai, Qianwei Wang, Haoxi Ran, Guanya Shi, Katia Sycara, and Yaqi Xie. pySpatial: Generating 3D visual programs for zero-shot spatial reasoning. *arXiv preprint arXiv:2603.00905*, 2026. URL https://arxiv.org/abs/2603.00905.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. SAT: Dynamic spatial aptitude training for multimodal language models. arXiv preprint arXiv:2412.07755, 2025. URL https://arxiv.org/abs/2412.07755v3. Version 3, revised December 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-MLLM: Boosting MLLM capabilities in visual-based spatial intelligence. arXiv preprint arXiv:2505.23747, 2025. URL https://arxiv.org/abs/2505.23747.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10632–10643, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_Thinking_in_Space_How_Multimodal_Large_Language_Models_See_Remember_CVPR_2025_paper.html.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. MMSI-Bench: A benchmark for multi-image spatial intelligence. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/fe5d4bd3e2af823701b5d8f70a3c7602-Abstract-Conference.html.

Yuncong Yang, Jiageng Liu, Zheyuan Zhang, Siyuan Zhou, Reuben Tan, Jianwei Yang, Yilun Du, and Chuang Gan. MindJourney: Test-time scaling with world models for spatial reasoning. arXiv preprint arXiv:2507.12508, 2025b. URL https://arxiv.org/abs/2507.12508.

Shoubin Yu, Yue Zhang, Zun Wang, Jaehong Yoon, Huaxiu Yao, Mingyu Ding, and Mohit Bansal. When and how much to imagine: Adaptive test-time scaling with world models for visual spatial reasoning. arXiv preprint arXiv:2602.08236, 2026. URL https://arxiv.org/abs/2602.08236.

Guibin Zhang, Yanwei Yue, Xiangguo Sun, Guancheng Wan, Miao Yu, Junfeng Fang, Kun Wang, Tianlong Chen, and Dawei Cheng. G-Designer: Architecting multi-agent communication topologies via graph neural networks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 76678–76692. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhang25cu.html.

Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. arXiv preprint arXiv:2503.14489, 2025. URL https://arxiv.org/abs/2503.14489.

Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. GPTSwarm: Language agents as optimizable graphs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 62743–62767. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhuge24a.html.