SPACE: Evidence Governance for Faithful Spatial Reasoning

Anonymous EMNLP submission

Abstract

Multimodal spatial reasoning often depends on intermediate beliefs about object identity, count, location, scale, and relations. These beliefs may appear coherent while remaining weakly grounded in the video, allowing unsupported spatial claims to influence the answer decision. We propose SPACE, a Spatial PlAnner and Claim-grounded Evidence governance framework for video spatial question answering. SPACE plans the evidence required by a question, collects structured local observations from video frames, assesses whether they support the required spatial claims, and applies an evidence-gated solver to produce an answer decision. By separating evidence acquisition from answer selection, SPACE makes intermediate spatial claims inspectable and constrains revision through grounded observations. Our analysis shows that reliable spatial reasoning depends on whether local evidence is supported by the cited frames, and that weakly grounded evidence can still lead to unreliable spatial decisions. These results suggest that faithful spatial reasoning requires not only correct answers, but also explicit governance over the spatial beliefs used to produce them.

1 Introduction

Spatial reasoning is a core capability for multimodal large language models (MLLMs). Unlike object recognition, spatial question answering often requires models to maintain beliefs about object identity, count, location, scale, temporal order, and relations across partial observations. This requirement is especially difficult in egocentric video and multi-view settings, where evidence is distributed across frames. In these settings, models must distinguish repeated views from distinct entities, align viewpoint-dependent directions, and integrate local observations into a coherent scene-level belief (Yang et al., 2024).

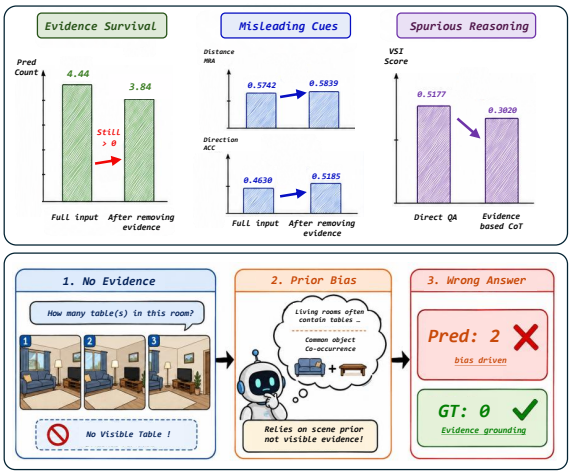

Figure 1: Cited evidence may not justify the underlying spatial reasoning. Evidence removal, misleading frames, and missing observations can lead to unsupported or prior-driven answers.

Existing benchmarks mainly evaluate final answer accuracy (Yang et al., 2024; Lin et al., 2026; Wang et al., 2025; Wu et al., 2025). However, a correct answer does not necessarily mean that the underlying spatial belief is visually grounded. A model may predict the right count while double-counting repeated views, choose the correct direction under an incorrect coordinate frame, or infer distance from object co-occurrence rather than supported comparison. As shown in Figure 1, cited evidence can be insufficient, misleading, or even reinforce spurious reasoning. Recent work has shown that spatial QA evaluation depends on whether the visual input contains enough evidence to support the answer (Zhang et al., 2026c; Xiao et al., 2026; Zhang et al., 2026a,b). We study a complementary question: when a model produces evidence or structured reasoning, are the spatial claims used to answer the question actually supported by the cited visual observations?

We refer to this failure mode as misgrounded spatial belief. A misgrounded belief is a plausi-

ble spatial claim that may even lead to a correct answer, but is not sufficiently supported by visual evidence. Such errors include claiming that an ambiguous object is visible, treating repeated observations as distinct instances, inferring relative direction from camera orientation rather than the question-defined frame, and judging routes or distances from landmark visibility without establishing connectivity or comparison. These failures are difficult to detect from answer accuracy alone because spatial answers compress richer latent commitments.

To address this problem, we propose SPACE, a Spatial PlAnner and Claim-grounded Evidence governance framework for multimodal spatial reasoning. Rather than treating spatial reasoning as free-form answer revision, SPACE organizes the process around three stages. The problem planner identifies the spatial task type, target entities, answer format, and evidence needed to verify the relevant spatial claims. The belief audit stage then collects structured local observations using evidence modules for common spatial requirements, including instance evidence for counting, candidate-anchor comparison for distance, coordinate-frame evidence for direction, route connectivity, and temporal appearance order. Finally, the revision solver maps the assessed evidence to an answer decision under an evidence-gated rule.

This design casts spatial reasoning as an auditable evidence process. The MLLM is used to collect local visual evidence, while answer selection is constrained by explicit support rules tied to the question. As a result, SPACE can indicate whether a decision is licensed by a supported new instance, a corrected coordinate frame, a better-supported spatial belief, or evidence that weakens a route claim. It also separates errors in evidence acquisition from errors in mapping evidence to an answer.

We further analyze evidence grounding at the claim level. Rather than treating natural-language rationales as faithful explanations, we examine whether local evidence claims are supported by their cited frames. Using detector-based diagnostics as an external proxy, we measure whether entity-level claims are visually supported and whether weakly grounded evidence is associated with lower decision reliability. Our analysis shows that structured evidence improves spatial QA only when it is grounded. Explicit solvers can reduce evidence-to-answer mapping errors, but unreliable local evidence remains a major bottleneck.

Our contributions are threefold. First, we formulate video spatial reasoning as an evidence-governed process, where answer-critical spatial claims must be verified against cited visual evidence rather than treated as implicit reasoning steps. Second, we introduce SPACE, a framework that plans evidence requirements, collects structured local evidence, and performs evidence-gated revision. Third, we provide claim-level diagnostics showing how evidence grounding affects revision reliability, offering a stricter view of faithful spatial reasoning beyond final answer accuracy.

2 Related Work

2.1 Spatial Reasoning

Spatial reasoning has become an important testbed for MLLMs because it requires models to form beliefs about scene structure and update these beliefs as visual evidence accumulates, rather than merely recognize isolated objects (Xiao et al., 2026; Yang et al., 2024). Recent efforts to improve spatial reasoning in MLLMs can be broadly grouped into two lines. One line focuses on test-time reasoning strategies, such as prompting, decomposition, cognitive-map construction, and inference-time search, which aim to improve spatial reasoning without changing model parameters (Yang et al., 2024, 2026; Hua et al., 2026). The other line improves spatial reasoning through post-training, using spatially grounded instruction data, preference optimization, or reinforcement learning to better align model behavior with spatial tasks (Ouyang et al., 2025; Chen et al., 2025; Yang et al., 2025; Wu et al., 2026). Together, these studies show that spatial reasoning remains difficult when the relevant evidence is partial, viewpoint-dependent, or distributed across observations. Recent work further argues that spatial evaluation should consider whether the supplied visual input is sufficient to answer the question (Zhang et al., 2026c). In this work, we study a complementary problem: whether the intermediate spatial evidence used to support an answer is grounded in the video, rather than only whether the final answer is correct.

2.2 Grounded Evidence and Multimodal Hallucination

A growing body of work studies whether MLLM outputs are grounded in visual input rather than

language priors or post-hoc explanations. This issue is closely related to multimodal hallucination, where models generate unsupported objects, attributes, relations, or reasoning steps (Guo et al., 2026; Li et al., 2025; Tu et al., 2025). Recent studies show that hallucination can arise within multimodal reasoning chains, and that longer reasoning may amplify weak visual cues into overconfident conclusions (Dong et al., 2026; Xu et al., 2026). Other work examines evidence grounding in multi-hop or multi-source multimodal QA, where answers must be linked to supporting text, tables, figures, or images (Xiang et al., 2026; Park et al., 2025). These studies show that answer correctness alone is insufficient for evaluating grounded reasoning. Spatial QA introduces an additional challenge: a cited frame may contain relevant objects but still fail to justify the required spatial claim, such as cross-view object identity, closest-object comparison, viewpoint-defined direction, or route connectivity (Xiao et al., 2026; Wang et al., 2025).

2.3 Evidence Auditing and Faithfulness Evaluation

Faithfulness evaluation asks whether explanations, rationales, or cited evidence actually support a model's prediction (Hu et al., 2025; Zhang et al., 2024). Existing grounding and hallucination methods can identify unsupported statements or check whether a response has visual support, but they often do not test whether the cited evidence is sufficient for the spatial claim being made (Guo et al., 2026; Zheng et al., 2025). Evidence-centric benchmarks move toward this goal through step-level annotations and source-level support (Wang et al., 2023), while ReVSI highlights the need to verify answerability from available visual observations (Zhang et al., 2026c). Our work addresses a complementary gap: even when a question is answerable from the video, the model's local evidence may not justify the spatial belief used to produce the final answer. We therefore propose an evidence audit that decomposes video spatial QA into evidence planning, structured evidence collection, and explicit evidence-to-answer revision.

3 Methodology

We introduce SPACE, a spatial evidence governance framework for video spatial question answering, as illustrated in Figure 2. Rather than directly prompting a multimodal model to revise its prediction, SPACE decomposes the question into auditable spatial requirements, gathers structured evidence from the video, and uses the audited evidence to govern answer selection. The goal is to address a common failure mode in spatial reasoning, where a model may produce a plausible prediction even when the underlying visual evidence does not support the spatial belief required by the question.

Given a video $V = \{f_1, \ldots, f_T\}$, a question q, and an initial prediction a_0, SPACE produces an evidence-grounded answer decision together with a diagnostic record of the evidence used during reasoning. The framework consists of three main stages, namely problem planning, belief auditing, and revision solving. In the planning stage, SPACE identifies the task type, registers key entities, and specifies the spatial belief and evidence requirements to be verified. The belief audit stage routes the query to task-relevant evidence types, acquires structured observations from the video, and assesses whether the current belief is sufficiently supported. The revision solver then aggregates the audited evidence, ranks candidate answers, and determines whether the initial prediction should be preserved or revised under the collected evidence.

3.1 Problem Planning

The problem planner maps the question and the initial prediction into a compact task specification $\mathcal{P}$. This specification makes explicit the spatial beliefs that must be verified before answer revision. It identifies the expected answer format, the entities and attributes involved, the spatial property under inspection, and the support rules used to evaluate the evidence. For example, a counting question requires instance-level support, a direction question requires a relation under a specified reference frame, and a navigation question requires evidence of feasible connectivity. In this way, the planner turns the original question into a set of answer-critical claims that can be audited.

Beyond identifying mentioned objects, the planner assigns functional roles to the entities and claims in the question. These roles indicate whether a claim is required for accepting an answer, auxiliary for resolving ambiguity, or invalidating when contradicted or unsupported. The planner also derives the reference condition needed for verification, such as a coordinate frame for direction, an anchor and criterion for compar-

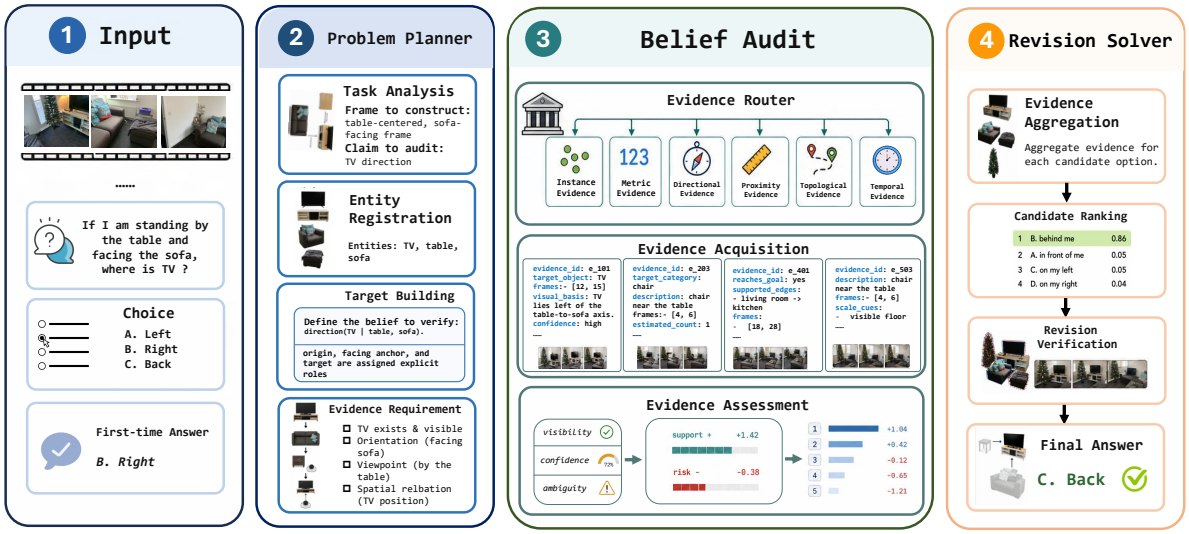

Figure 2: Overview of SPACE. The framework plans the target spatial belief, audits claim-specific visual evidence, and revises the initial answer using the grounded evidence.

ison, or a connectivity constraint for navigation. These decisions determine how evidence should be acquired and assessed in later stages, so that the system looks for observations that can justify the intended claim rather than frames that only contain the queried objects.

The same specification records invalid support patterns that commonly lead to unreliable revisions. These include treating object visibility as relation support, confusing camera-left with egocentric left, relying only on perspective cues for distance, or accepting a fixed answer without visual justification. By making these failure modes explicit, the planner constrains subsequent evidence collection and provides the claim-level rules used by the assessment stage.

3.2 Belief Audit

The task specification from the previous stage serves as the interface between planning and evidence acquisition. Let $V = \{f_t\}_{t=1}^{T}$ denote the sampled video frames. Given $\mathcal{P}$, SPACE selects a set of audit modules from an evidence-module library according to the spatial belief to be verified. We denote the selected modules as

$$\mathcal{M} = R(\mathcal{P}, \mathbb{M}), \tag{1}$$

where $\mathbb{M}$ is the module library and R is the evidence router. Each selected module provides an instruction fragment and an output schema for collecting a specific type of supporting observation. The router composes these modules with $\mathcal{P}$ and V to construct the audit prompt. The MLLM is then asked to populate the evidence schema, so that the audit produces structured claim-level observations rather than a direct answer revision.

The module library defines evidence interfaces for different spatial claims. For example, directional modules require the reference frame under which a relation is evaluated, comparison modules require the anchor and selection criterion, and navigation modules require evidence of feasible or blocked transitions. Each module also specifies the frame-level support to cite and the uncertainty or failure conditions to record. The resulting audit record therefore contains claim-specific evidence rather than a generic description of visible objects. This design treats object presence as a prerequisite for spatial evidence, but not as sufficient support for the spatial belief under audit.

3.3 Evidence Assessment

After evidence acquisition, SPACE does not ask the model to produce another free-form reasoning chain. Instead, it asks a more constrained question: what is licensed by the collected evidence? The assessment stage converts heterogeneous audit records into compact support signals for the revision solver.

For numeric questions, including counting and estimation, the collected evidence is converted into an evidence-conditioned estimate. This estimate is derived from structured observations such as supported instances, valid comparison cues, visual scale cues, or layout cues, depending on the planned evidence requirement. It is paired with support indicators that record whether the initial prediction is visually supported, whether a differ-

ent value is justified, and whether the evidence is strong, weak, or uncertain.

Different from numeric questions, non-numeric spatial questions usually correspond to relational or state-like beliefs rather than scalar estimates. Their evidence therefore cannot be aggregated by simply counting instances or averaging measurements. SPACE addresses this by reformulating assessment as candidate belief verification. Each candidate belief is linked to the evidence items relevant to its required spatial claim, and these items are checked against the support rules in the problem plan and evidence schema. Evidence that satisfies a required condition adds support, partial or ambiguous evidence adds weak support, and contradictory or invalid evidence reduces support. We summarize this aggregation as

$$s_{\text{belief}}(b) = \sum_{e \in \mathcal{E}_b} \phi(e; \mathcal{P}),$$

where $\mathcal{E}_b$ is the evidence associated with belief b, $\mathcal{P}$ is the problem plan, and ϕ is a deterministic rule that maps each structured evidence item to its support contribution. The rule is instantiated by the evidence type, rewarding claim-specific support such as valid visibility, reliable reference conditions, consistent relations, or supported connectivity, while penalizing missing, ambiguous, contradicted, or insufficient evidence.

The resulting estimate or support score summarizes what the evidence supports. It is not treated as a calibrated probability, and it does not directly decide whether the initial prediction should be preserved or revised. That decision is made by the revision solver under the evidence-gated rule described next.

3.4 Revision Solver

The revision solver maps the assessed evidence to an answer decision without introducing another free-form reasoning step. Its role is to close the evidence governance loop: the prediction is selected according to what the audited evidence licenses, rather than according to a newly generated explanation.

The solver follows an evidence-gated principle. An alternative prediction is adopted only when it receives stronger support from the audited evidence and the initial prediction is weakened, contradicted, or left unsupported. When the evidence is incomplete or ambiguous, SPACE avoids treating plausibility as sufficient grounds for revision. This prevents self-correction from becoming an unconstrained rewriting process.

The output of the solver is therefore not only an answer decision, but also a traceable evidence judgment. Because the decision is derived from the planned requirements, structured observations, and support assessment, each revision can be attributed to the visual evidence that licenses it.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate SPACE on VSI-Bench (Yang et al., 2024) and OST-Bench (Lin et al., 2026). VSI-Bench covers spatial QA in egocentric indoor videos, including counting, distance, scale, direction, route, and temporal order questions. OST-Bench evaluates online spatiotemporal understanding for embodied agents, where answers depend on incrementally observed context. Together, they test both offline video spatial reasoning and memory-dependent online spatial reasoning.

Base models. We evaluate SPACE across different model scales and families. On VSI-Bench, we use Qwen3-VL-32B, Qwen3-VL-8B (Bai et al., 2025), and MiMo-VL-7B (Team et al., 2025) as representative MLLM base models. We further include SenseNova-U1-8B-MoT (Diao et al., 2026), a recent multimodal reasoning model with strong spatial reasoning performance in our evaluation. On OST-Bench, we use Qwen3-VL-8B and SenseNova-U1-8B-MoT to compare a compact open model with a spatially strong model under online spatial reasoning.

Baselines. Following Hua et al. (2026), we compare SPACE with Direct, Chain-of-Thought (CoT) (Wei et al., 2022), Tree-of-Thoughts (ToT) (Yao et al., 2023), Least-to-Most prompting (LtM) (Zhou et al., 2023), Cognitive Map prompting (CM) (Yang et al., 2024), and TRACE (Hua et al., 2026). These baselines cover both general reasoning prompts and spatially structured prompting methods. In contrast, SPACE explicitly audits whether answer-critical spatial claims are supported by visual evidence before producing the final answer.

Evaluation metrics. Following prior work Yang et al. (2024); Hua et al. (2026), we report accuracy for multiple-choice questions. For numerical questions, we report mean relative accuracy (MRA),

Models	Methods	Avg.	Numerical Answer				Multiple-Choice Answer			
			Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route	Order
Qwen3-VL-32B	Direct	57.61	56.69	52.24	66.91	66.90	60.32	57.01	35.09	**65.72**
	CoT	54.50	56.33	54.15	67.20	68.70	50.32	54.45	29.82	55.03
	ToT	59.84	**57.97**	**55.39**	**68.47**	68.50	**61.29**	65.71	39.47	61.95
	LtM	58.48	57.01	54.38	67.68	68.20	58.39	61.78	35.96	64.47
	CM	56.68	56.41	53.76	68.16	68.05	54.52	59.42	32.46	60.69
	TRACE	59.53	56.61	54.79	68.30	**68.80**	60.65	66.49	37.72	62.89
	SPACE (Ours)	**60.37**	57.65	54.56	67.76	66.35	60.97	**69.90**	**40.35**	65.41
Qwen3-VL-8B	Direct	51.77	52.62	43.25	63.80	60.50	**57.42**	44.86	33.33	58.18
	CoT	30.20	24.66	21.47	15.78	11.85	45.81	38.22	27.19	56.60
	ToT	52.10	53.07	44.45	64.19	60.70	56.13	50.26	34.21	53.77
	LtM	50.58	52.27	43.87	63.97	60.55	55.16	47.64	32.46	48.74
	CM	49.01	51.47	42.86	63.80	60.50	53.23	45.55	31.58	43.08
	TRACE	53.31	52.55	45.09	64.50	60.70	56.45	52.09	**35.96**	**59.12**
	SPACE (Ours)	**54.90**	**53.90**	**48.20**	**67.85**	**61.50**	56.13	**58.38**	35.09	58.18
MiMo-VL-7B	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
	TRACE	41.42	33.27	31.51	58.67	41.56	39.44	35.33	28.87	51.29
	SPACE (Ours)	**44.16**	**39.20**	**36.59**	**62.38**	**45.15**	40.65	**43.72**	32.46	**53.14**
SenseNova-U1-8B-MoT	Direct	57.72	**61.75**	44.98	63.37	**63.80**	**51.29**	59.42	42.98	**74.21**
	CoT	48.44	41.59	36.04	55.24	55.55	42.90	54.71	38.60	62.89
	ToT	53.84	53.15	45.83	63.57	62.05	50.32	56.81	38.60	60.38
	LtM	52.64	52.47	45.32	63.37	61.85	49.35	54.71	37.72	56.29
	CM	50.56	50.76	44.84	61.44	59.85	48.06	52.88	35.09	51.57
	TRACE	54.30	52.63	46.31	63.80	62.05	50.32	57.33	40.35	61.64
	SPACE (Ours)	**59.52**	61.00	**50.69**	**67.73**	63.00	50.65	**63.87**	**45.61**	73.58

Table 1: Evaluation results on VSI-Bench. We report average and task-level scores across numerical-answer and multiple-choice spatial reasoning tasks. Best results are shown in bold.

which measures whether the predicted value falls within relative-error thresholds. We report both overall average scores and task-level results.

4.2 Main Results

Tables 1 and 2 report the main results. On VSI-Bench, SPACE achieves the best average performance across all four base models and consistently outperforms TRACE. This shows that its gains do not simply come from adding an intermediate textual spatial representation. Instead, the advantage comes from explicitly checking whether the spatial claims used for revision are supported by the cited visual evidence.

The improvements are most evident on questions where seeing the relevant objects is not enough to justify the answer. Relative direction improves across all base models, and route planning also benefits on the stronger base models. These tasks require the model to verify coordinate frames, spatial relations, and feasible transitions, rather than merely identify object co-occurrence in a frame. SPACE is therefore particularly useful when a plausible spatial statement must be distinguished from a visually grounded one.

For numerical-answer tasks, the pattern is more selective. SPACE improves object counting on most base models and brings consistent gains on absolute distance and object size. These tasks benefit from filtering weakly supported estimates and checking whether the cited observations actually justify the predicted quantity. Room-size estimation is less stable, likely because sparse egocentric views often provide limited metric cues. This suggests that SPACE is most effective when the necessary evidence is present but must be selected, compared, and validated.

On OST-Bench, SPACE also improves both evaluated base models. The gains are concentrated in agent-object spatial relationship questions, especially direction and distance, where the answer depends on observations accumulated over time rather than the current view alone. The smaller changes on directly visible information and some agent-state categories lead to a consistent interpretation. Evidence governance helps most when

Models	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.
Qwen3-VL-8B	Direct	49.40	**83.99**	56.22	66.98	80.15	61.82	41.00	24.57	19.08	**45.08**	41.51	36.21	59.04	41.37	49.90	35.52
	TRACE	51.57	83.22	57.71	**67.62**	**81.27**	63.18	48.54	28.99	23.66	41.03	44.23	39.47	60.89	44.79	**51.78**	39.25
	SPACE (Ours)	**52.98**	82.60	**58.84**	66.35	79.03	**63.64**	**52.03**	**31.94**	**28.75**	44.37	**46.54**	**43.10**	**62.36**	**47.07**	48.64	**42.33**
SenseNova-U1-8B-MoT	Direct	54.27	**86.37**	60.92	66.35	77.15	**67.50**	48.20	27.52	20.10	48.04	45.07	40.94	59.04	44.16	**51.78**	37.80
	TRACE	52.95	85.07	60.33	64.44	80.15	65.91	47.08	26.54	16.03	46.01	41.51	40.58	62.36	41.60	48.85	37.36
	SPACE (Ours)	**57.97**	84.53	**65.23**	**67.62**	**80.52**	61.36	**54.49**	**39.56**	**32.32**	**51.49**	**48.43**	**47.46**	**66.79**	**44.44**	51.15	**46.15**

Table 2: Evaluation results on OST-Bench. We report average and task-level scores across agent visible information, agent-object spatial relationships, and agent state. Best results are shown in bold.

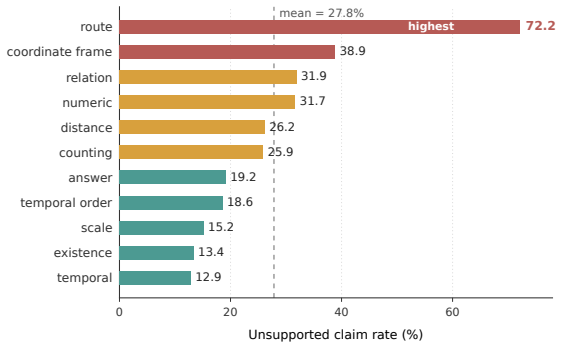

Figure 3: Unsupported claim rates across spatial claim types.

Task	USCR ↓	EPR ↑
Counting	3.8	73.7
Temporal	15.0	73.0
Obj. Distance	18.2	76.3
Room Size	21.7	68.0
Obj. Size	24.1	58.6
Rel. Distance	26.2	61.3
Rel. Direction	32.1	52.4
Route	50.8	5.3

Table 3: Task-level evidence reliability for SPACE. USCR measures the fraction of unsupported answer-critical claims, and EPR measures the fraction of questions whose answer-critical claims all pass the audit.

spatial reasoning requires deciding which observations constitute valid support for the belief being revised.

4.3 Evidence Grounding Analysis

The task-level results show that the gains of SPACE are concentrated in questions requiring spatial relations, reference frames, distances, or transitions. Since these questions depend on the validity of intermediate spatial claims, we further examine whether such claims are grounded in the cited video frames. We perform this analysis on VSI-Bench using Qwen3-VL-8B, and relate the grounding behavior to the reliability of the evidence used for answer revision.

Unsupported spatial claims. We report the unsupported rate for each claim category, defined as the fraction of audited claims labeled unsupported or contradicted. Figure 3 shows that grounding failures are not uniform across spatial claims. Route claims have the highest unsupported rate of 72.2%, followed by coordinate-frame, relation, numerical, distance, and counting claims, whereas existence and temporal claims are much better grounded. This contrast suggests that unsupported evidence is concentrated in claims that require more than local object visibility. Route and relation claims must establish connectivity, reference frames, or object-to-object relations across observations, while numerical and distance claims often rely on weak metric cues. Existence claims are easier to ground because they can often be verified from a single visible instance.

Entity and frame grounding. We further examine grounding at the entity and frame levels. Among 6,737 entity-level claims, 4,891 can be matched to the SAM3 detector (Carion et al., 2026), yielding a vocabulary match rate of 72.6%. This suggests that most evidence claims refer to plausible visual entities. The harder question is whether those entities are actually supported in the cited frames. We find that 67.4% of the claims have valid cited-frame support, whereas 22.6% overstate object visibility by relying on absent, partial, or ambiguous visual evidence. This gap indicates that entity-level plausibility does not necessarily translate into frame-grounded support for the spatial claim.

Decision grounding. We further measure whether the final decision is internally grounded in the solver's diagnostic evidence state. For each example, we construct an answer-level claim from the final decision record and count it as grounded when the corresponding evidence support and solver confidence are both at least medium. Under this criterion, 1,908 out of 2,362 decisions are grounded, yielding a decision grounding rate of

Figure 4: Qualitative example of YAML-based evidence auditing. The baseline undercounts by localizing only two plants, while SPACE records three visible instances and revises the answer through `count_reconciliation`.

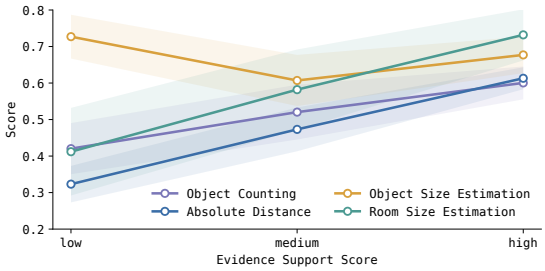

Figure 5: Downstream performance under low-, medium-, and high-ESS regimes across numerical spatial reasoning tasks.

80.8%. This metric is not an external correctness measure; rather, it diagnoses whether the solver's decision is backed by the evidence state it receives. The result suggests that most decisions are accompanied by sufficient internal support signals, while the remaining cases usually reflect incomplete, ambiguous, or weakly grounded evidence.

Task-level reliability. We then examine how evidence reliability varies across tasks. Table 3 reports two diagnostics at different levels. USCR measures the fraction of unsupported or contradicted answer-critical claims, while Evidence Pass Rate measures whether the overall evidence record satisfies the planned support requirements. Counting has the lowest USCR at 3.8%, suggesting that instance-level evidence is comparatively easy to verify. In contrast, route planning has a USCR of 50.8% and an Evidence Pass Rate of only 5.3%. This gap reflects the difference between local visual support and compositional spatial support. Route questions require connected and feasible transitions, so a missing key transition can invalidate the whole evidence record.

Evidence support and downstream performance. We finally examine whether stronger evidence support leads to better downstream answer quality on numeric tasks, whose outputs can be directly scored. For each prediction, we compute an Evidence Support Score (ESS) by averaging the auditor support scores of its answer-critical claims, and group examples into low-, medium-, and high-support regimes. Figure 5 shows that higher ESS generally improves performance, especially for counting, absolute distance, and room-size estimation. Object-size estimation is an exception, likely due to strong category- and prior-driven size biases, as also observed by Wu et al. (2025). These results suggest that evidence grounding often improves reliability, but accurate numerical prediction may still require better calibration beyond visual support alone.

4.4 Qualitative Analysis

Figure 4 presents a representative counting example. The baseline predicts two plants after grounding the answer on only two visible instances. SPACE instead collects structured instance-level evidence and identifies three spatially distinct plants: one near the wooden chest, one behind the TV, and one behind the keyboard. The `count_reconciliation` solver revises the answer from two to three. This case illustrates how evidence governance improves counting by maintaining auditable object-instance beliefs rather than relying on free-form rationalization.

5 Conclusion

This paper identifies a citation-validity gap in multimodal spatial reasoning: cited evidence may be visually relevant yet fail to support the spatial facts required by the answer. We attribute this gap to claim-to-evidence drift and coverage mismatch, and propose a claim-grounded evidence audit that decomposes provisional answers into typed spatial claims and verifies whether the cited visual evidence is sufficient for each claim. Video spatial QA provides a natural testbed, where frame-level evidence makes the claim-observation support relation directly inspectable.

6 Limitations and Future Work

SPACE improves the auditability of video spatial reasoning, but its performance is still limited by the quality of local evidence acquisition. When relevant objects are missed, only partially visible, or observed from ambiguous viewpoints, the resulting evidence record may remain incomplete. This limitation is especially important for counting and relational questions, where the system must distinguish distinct instances from repeated views and verify that the cited frames support the intended spatial claim, rather than merely contain the relevant objects. The counting example in Figure 4 illustrates the value of instance-level evidence, but it also shows that such corrections depend on whether missing instances can be recovered from the video.

Confidence and numerical calibration also remain challenging. Although the results suggest that answer confidence generally increases with evidence confidence, some cases remain overconfident despite weak claim-level support. For numeric questions, especially object-size estimation, the model may also be affected by category- and prior-driven biases. In such cases, even visually grounded evidence may not be sufficient to produce a well-calibrated numerical estimate. These errors indicate that the solver may still assign stronger confidence than the available visual evidence warrants, especially when the evidence is incomplete, ambiguous, or mediated by strong priors.

Finally, the current module library covers a limited set of spatial claim types, and multi-stage auditing introduces additional inference cost. Future work should make evidence acquisition more adaptive, extend the evidence modules to richer spatial phenomena such as occlusion, fine-grained geometry, and route connectivity, and better align answer confidence and numerical estimates with the strength of visual support.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, and 19 others. 2026. SAM 3: Segment anything with concepts. In *The Fourteenth International Conference on Learning Representations*.

Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Xiang An, Yan Feng, Peng Pei, Xunliang Cai, and 1 others. 2025. Think with 3d: Geometric imagination grounded spatial reasoning from limited views. *arXiv preprint arXiv:2510.18632*.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, Yan Li, Yubo Wang, Zhijie Cao, Zhiqian Lin, Zhitao Yang, Zhongang Cai, Yuwei Niu, Yue Zhu, Bo Liu, and 39 others. 2026. Sensenova-u1: Unifying multimodal understanding and generation with neo-unify architecture. *Preprint*, arXiv:2605.12500.

Bowen Dong, Minheng Ni, Zitong Huang, Guanglei Yang, Wangmeng Zuo, and Lei Zhang. 2026. Mirage: Assessing hallucination in multimodal reasoning chains of mllm. *Advances in Neural Information Processing Systems*, 38:122910–122955.

Pinxue Guo, Chongruo Wu, Xinyu Zhou, Lingyi Hong, Zhaoyu Chen, Jinglun Li, Kaixun Jiang, Sen-Ching Samson Cheung, Wei Zhang, and Wenqiang Zhang. 2026. Seeing is believing: Rich-context hallucination detection for mllms via backward visual grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 30871–30879.

Nan Hu, Jiaoyan Chen, Yike Wu, Guilin Qi, Hongru Wang, Sheng Bi, Yongrui Chen, Tongtong Wu, and Jeff Z Pan. 2025. Can llms evaluate complex attribution in qa? automatic benchmarking using knowledge graphs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17096–17118.

Jiacheng Hua, Yishu Yin, Yuhang Wu, Tai Wang, Yifei Huang, and Miao Liu. 2026. Unleashing spatial reasoning in multimodal large language models via textual representation guided reasoning. *arXiv preprint arXiv:2603.23404*.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. 2025. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13723–13733.

Jingli Lin, Chenming Zhu, Runsen Xu, Xiaohan Mao, Xihui Liu, Tai Wang, and Jiangmiao Pang. 2026. Ost-bench: Evaluating the capabilities of mllms in online spatio-temporal scene understanding. *Advances in Neural Information Processing Systems*, 38.

Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. 2025. SpaceR: Reinforcing MLLMs in video spatial reasoning. *Preprint*, arXiv:2504.01805.

Jiwon Park, Seohyun Pyeon, Jinwoo Kim, Rina Carines Cabal, Yihao Ding, and Soyeon Caren Han. 2025. Dochop-qa: Towards multi-hop reasoning over multimodal document collections. *arXiv preprint arXiv:2508.15851*.

Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, and 55 others. 2025. Mimo-vl technical report. *Preprint*, arXiv:2506.03569.

Yahan Tu, Rui Hu, and Jitao Sang. 2025. Ode: Open-set evaluation of hallucinations in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19836–19845.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and 1 others. 2023. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, and 1 others. 2025. Mindcube: Spatial mental modeling from limited views. *arXiv e-prints*, pages arXiv–2506.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. 2026. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *Advances in Neural Information Processing Systems*, 38:143297–143330.

Mingrui Wu, Zhaozhi Wang, Fangjinhua Wang, Jiaolong Yang, Marc Pollefeys, and Tong Zhang. 2025. From indoor to open world: Revealing the spatial reasoning gap in mllms. *arXiv preprint arXiv:2512.19683*.

Biao Xiang, Soyeon Caren Han, and Yihao Ding. 2026. Bridge: Benchmark for multi-hop reasoning in long multimodal documents with grounded evidence. *arXiv preprint arXiv:2603.07931*.

Yuxi Xiao, Longfei Li, Shen Yan, Xinhang Liu, Sida Peng, Yunchao Wei, Xiaowei Zhou, and Bingyi Kang. 2026. Spatialtree: How spatial abilities branch out in mllms. In *The First Workshop on Efficient Spatial Reasoning*.

Zhongxing Xu, Chengzhi Liu, Qingyue Wei, Juncheng Wu, James Zou, Xin Wang, Yuyin Zhou, and Sheng Liu. 2026. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *Advances in Neural Information Processing Systems*, 38:82878–82905.

Jianing Yang, Shunyu Yang, Anirudh W. Gupta, Runnan Han, Li Fei-Fei, and Saining Xie. 2024. Thinking in space: How multimodal large language models see, remember, and recall spaces. *Preprint*, arXiv:2412.14171.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, and 1 others. 2025. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*.

Yuncong Yang, Jiageng Liu, Zheyuan Zhang, Siyuan Zhou, Reuben Tan, Jianwei Yang, Yilun Du, and Chuang Gan. 2026. Mindjourney: Test-time scaling with world models for spatial reasoning. *Advances in Neural Information Processing Systems*, 38:109855–109885.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Pingyue Zhang, Zihan Huang, Yue Wang, Jieyu Zhang, Letian Xue, Zihan Wang, Qineng Wang, Keshigeyan Chandrasegaran, Ruohan Zhang, Yejin Choi, and 1 others. 2026a. Theory of space: Can foundation models construct spatial beliefs through active exploration? In *The Fourteenth International Conference on Learning Representations*.

Weijia Zhang, Mohammad Aliannejadi, Yifei Yuan, Jiahuan Pei, Jia-Hong Huang, and Evangelos Kanoulas. 2024. Towards fine-grained citation evaluation in generated text: A comparative analysis of faithfulness metrics. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 427–439.

Yi Zhang, Youya Xia, Yong Wang, Meng Song, Xin Wu, Wenjun Wan, Bingbing Liu, AiXue Ye, Hongbo Zhang, and Feng Wen. 2026b. Ssr: Pushing the limit of spatial intelligence with structured scene reasoning. *arXiv preprint arXiv:2603.00409*.

Yiming Zhang, Jiacheng Chen, Jiaqi Tan, Yongsen Mao, Wenhu Chen, and Angel X. Chang. 2026c. Revsi: Rebuilding visual spatial intelligence evaluation for accurate assessment of vlm 3d reasoning. *arXiv preprint arXiv:2604.24300*.

Kening Zheng, Junkai Chen, Yibo Yan, Xin Zou, Huiyu Zhou, and Xuming Hu. 2025. Reefknot: A comprehensive benchmark for relation hallucination evaluation, analysis and mitigation in multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6193–6212.

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*.

Appendix

This appendix provides technical details of our prompting templates and deterministic scoring rules. Appendix A presents the problem planner, evidence collection, and decision record templates, while Appendix A.2 describes the candidate scoring and revision rules used by SPACE.

A Prompting and Scoring Details

A.1 SPACE Prompt and Record Templates

SPACE separates video spatial QA into three stages: problem planning, evidence collection, and evidence-to-answer decision. The first stage converts the raw question and the initial answer into a structured problem plan. The second stage asks the MLLM to populate task-specific evidence schemas rather than directly revise the answer. The final stage applies deterministic evidence-to-answer rules and returns a diagnostic decision record.

Problem Planner Prompt

Input:
```
question: {question}
question_type: {question_type}
choices: {A: choice text, B: choice text, ...}
initial_answer: {a_0}
```

Output fields:
```
problem_plan:
  task_family: {primary, subtype}
  answer_format: {type, must_output_option_letter, choices}
  entities: {target_entities, anchor_objects, candidate_objects}
  question_semantics: {asks_for, target, reference, quantity, answer_space}
  evidence_requirements:
    required_observations: [{id, description, module, target, scope}]
    disambiguation_checks: [duplicate identity, occlusion, scale cue, ...]
    invalid_support_patterns: [visibility_only, perspective_only, ...]
  reasoning_plan: [{step, uses, output}]
  evidence_modules: {required: [module ids], optional: []}
  solver_strategy: {type, allow_revision, keep_fixed_answer_only_if}
```

Evidence Collection Prompt

```
You are an Evidence Collector for egocentric indoor spatial QA.
Do not answer the question. Do not decide keep/revise. Collect evidence only.
Use the ProblemPlan and selected Evidence Bank modules below.
Pay special attention to question semantics, evidence requirements, and reasoning plan.
Do not treat mere visibility as support unless the question asks for visibility.
For multiple-choice questions, compare all candidates under the same evidence standard.
If choices are present, keep option letters as candidate IDs.
Return STRICT YAML only.
```

ProblemPlan:
```
{serialized problem plan}
```

Selected module instructions:
```
[candidate_option_parser] Parse all answer choices exactly.
[entity_grounding] Record visibility for targets, anchors, and candidates.
[counting_evidence] List distinct visible instances and deduplicate repeated views.
[numeric_estimate] Estimate value, range, scale cues, and fixed-answer support.
[distance_relation] Compare all candidates against the same anchor.
[direction_relation] Define the coordinate frame and target direction.
[coordinate_frame_check] Verify origin, facing anchor, and frame reliability.
[route_topology] Check route connectivity, goal reachability, and blockage.
[appearance_order] Determine first visible frame/window for each category.
```

Required output shape:
```
collected_evidence:
  modules_used: [module ids]
  evidence: {schema fields required by selected modules}
```

Decision Record Prompt

Input:
```
ProblemPlan, collected_evidence, initial_answer
```

Optional candidate ranking:
```
{task_specific_ranking}:
  best_candidate: {option letter}
  best_score: {score}
  scores: {A: {score, sources, reasons}, ...}
```

Final decision:

```
final_decision:
  original_answer: {a_0}
  final_answer: {a*}
  answer_to_use: {a*}
  revision_decision: {keep | revise | uncertain_keep | uncertain_revise}
  evidence_support_level: {strong | medium | weak | none | uncertain}
  confidence: {high | medium | low}
  revision_reason: {evidence-based rationale}
```

A.2 Scoring and Revision Rules

The solver consumes the structured evidence record produced by the Evidence Bank. Different task families expose different evidence fields, so SPACE does not use one universal probability model. Instead, it applies field-based deterministic rules matched to the solver strategy specified in the problem plan. Candidate-ranking tasks produce scores over answer options, while counting and numeric tasks produce evidence-conditioned estimates.

For tasks with candidate-level evidence, the solver assigns each candidate a deterministic evidence score:

$$s_k = \sum_{e \in \mathcal{E}_k} \phi(e; \mathcal{P}),$$

where $\mathcal{E}_k$ denotes evidence associated with candidate o_k, and ϕ is a task-conditioned rule over the structured evidence fields. The score is used only to rank candidates within the same question and is not interpreted as a calibrated probability.

The shared support-strength mapping used by candidate-ranking rules is:

$$R_{\text{str}} = \begin{cases} 3 & \texttt{direct} \text{ or } \texttt{strong}, \\ 2 & \texttt{indirect}, \\ 1 & \texttt{weak}, \\ 0 & \texttt{entity_only} \text{ or } \texttt{invalid}. \end{cases}$$

For candidate-ranking tasks, let a_0 be the initial answer and $k^* = \arg\max_k s_k$. SPACE revises the answer when the best-supported candidate differs from the initial answer and exceeds the task-specific support threshold:

$$a^* = \begin{cases} \ell_{k^*}, & \ell_{k^*} \neq a_0 \ \wedge \ s_{k^*} \geq \tau, \\ a_0, & \text{otherwise}. \end{cases}$$

In our implementation, the default threshold for candidate-ranking overrides is $\tau = 3$. Counting and numeric tasks do not use this candidate-ranking rule. They revise according to the evidence-conditioned estimate and its support fields, such as `fixed_answer_supported`, `value_range`, `visible_instances`, and `confidence`. The final output includes the problem plan, selected modules, structured evidence, cited frames, candidate rankings when available, revision decision, and solver rationale.

Task	Support fields	Revision gate	Failure signals
Counting	instance, identity, frames	distinct count	duplicate, non-target, missing
Numeric	target, scale, range	supported value	no scale, low conf., out of range
Rel. distance	anchor, candidate, layout	best-supported candidate	missing anchor, perspective only
Rel. direction	anchor, axis, same frame	frame-aligned direction	unclear frame, direction mismatch
Route	landmark, edge, path	reachable goal	blocked, missing edge
Appearance order	category, first view, order	option-matched order	missing category, unclear first view
Generic choice	candidate, claim, frames	claim-grounded option	visibility only, partial support

Table 4: Compact summary of the evidence-to-revision rules in SPACE. Support fields increase the score of a candidate or estimate. Revision gates specify the minimum evidence needed to change the initial answer. Failure signals weaken the candidate or block revision when the cited frames do not support the required spatial claim.