VERIFICATION BEFORE COMMITMENT: ACTIVE ENVIRONMENTAL VERIFICATION FOR EMBODIED UNDERSTANDING

ABSTRACT

Embodied agents must often infer environmental states from partial observations. However, limited evidence may already yield a plausible interpretation without being sufficient to establish its correctness, leading agents to commit prematurely. We address this gap between interpretation plausibility and evidence sufficiency through a verification-based policy, which treats a proposed answer as a provisional hypothesis and shifts subsequent interaction from answer seeking to active hypothesis testing. Using embodied question answering (EQA) as a testbed, we introduce Sure-EQA, a training-free framework that re-examines existing visual support from alternative viewpoints and probes unseen regions for answer-changing evidence. Verification findings then guide whether the candidate should be accepted, further tested, or revised. Experiments on A-EQA and EXPRESS-Bench show consistent improvements in answer correctness and path-weighted performance. Under matched additional observation, Sure-EQA also outperforms continued exploration in reducing missing evidence and correcting mistaken interpretations.

1 INTRODUCTION

Embodied agents seek to understand their surroundings and act in the environment to accomplish a wide range of tasks. As they move and observe, agents gradually accumulate information about objects, scenes, and their relationships (Szot et al., 2025; Yang et al., 2025a). Recent advances in vision-language models (VLMs) have brought strong visual understanding and semantic reasoning capabilities to embodied systems (Deitke et al., 2025; Xu et al., 2025). By interpreting observations gathered during exploration and relating them to knowledge learned through large-scale pretraining, VLM-based agents can form richer scene interpretations and better guide subsequent exploration (Ren et al., 2024; Yang et al., 2025b; Jiang et al., 2025). Yet the ability to form plausible interpretations from partial observations does not guarantee that sufficient environmental evidence has been gathered to support them.

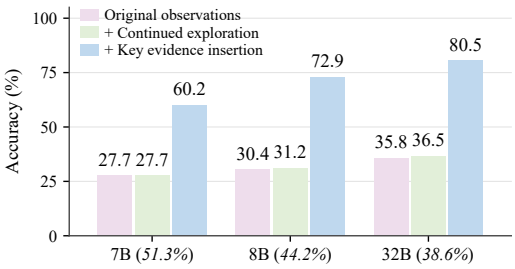


Figure 1: Diagnostic analysis of embodied agents using VLM backbones of different sizes. Parentheses report full-dataset missing-evidence rates. Bars compare fully correct answer rates on each agent's evidence-missing subset using original observations, continued exploration, or manually inserted key evidence.

This can lead embodied agents to commit prematurely to an interpretation that appears plausible but remains insufficiently verified. A VLM can combine limited observations with its learned knowledge to fill in what has not been directly observed, making an incomplete understanding appear complete (Luo et al., 2024; Wu et al., 2026). For example, when asked "What color are the dining room walls?", an agent may observe a light-blue wall together with a bowl and infer that the observed room is the dining room, leading to the answer "light blue". The wall color is directly observed, but the room identity is only inferred. A broader view may instead reveal that the wall belongs to a bedroom. To examine whether this problem occurs systematically, we conduct a diagnostic study on EXPRESS-Bench (Jiang et al., 2025). At each proposed stop, we examine whether the observations the agent deems sufficient to answer contain the annotated visual evidence supporting the ground-truth answer. As shown in Figure 1, key evidence is frequently still missing when

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.

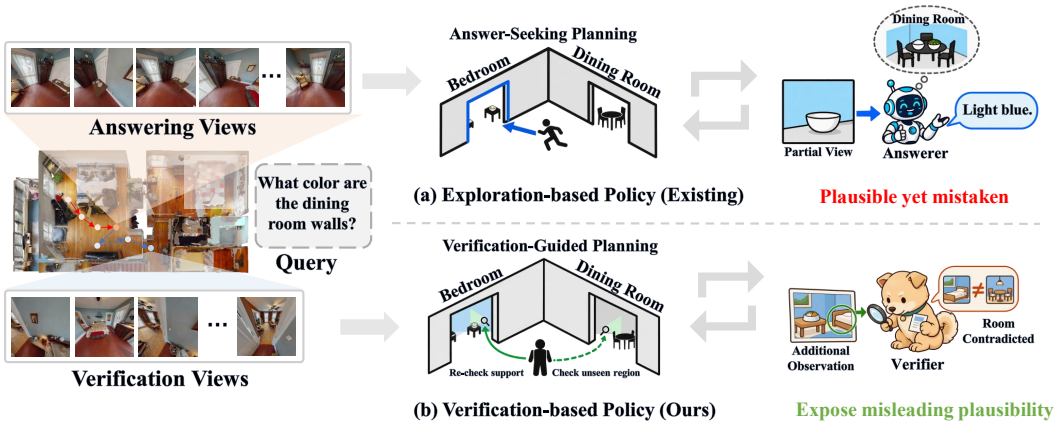


Figure 2: From answer seeking to hypothesis testing. (a) An exploration-based policy seeks an answer. (b) A verification-based policy redirects subsequent interaction to test the proposed answer against the environment.

agents propose to stop. Manually supplying this evidence substantially improves answer correctness, whereas simply continuing the original exploration yields only limited gains. These results reveal a gap between interpretation plausibility and evidence sufficiency. Agents are often capable of revising their answers when key evidence becomes available, yet their existing exploration policy may fail to acquire such evidence before commitment.

The difficulty is that a plausible interpretation can itself become a reason to stop acquiring the evidence needed to verify it. Once partial observations and learned knowledge yield a plausible answer, an answer-seeking agent following an exploration-based policy may regard the task as resolved, leaving the interpretation's underlying assumptions untested. We therefore advocate a shift in the agent's interaction objective once a candidate answer is formed, from answer seeking to active hypothesis testing (Figure 2). Rather than treating the proposed answer as a conclusion ready for acceptance, a verification-based policy treats it as a provisional hypothesis to be tested through further interaction with the environment. Subsequent interaction is directed toward examining the evidential basis of the proposed interpretation and seeking evidence that may support, challenge, or revise it.

To investigate this verification-based policy, we adopt embodied question answering (EQA) as a testbed, where agents can actively acquire new observations and express their current understanding through answers to questions (Majumdar et al., 2024; Jiang et al., 2025). These answers make the agent's interpretation explicit, enabling verification against subsequently acquired environmental evidence. Accordingly, we introduce Sure-EQA, a training-free framework that performs two complementary types of verification. Evidence verification re-examines visual support for the candidate from alternative viewpoints to assess whether the original interpretation remains valid, while coverage verification probes unexplored regions for potentially answer-changing evidence that may have been missed. Within each verification round, accumulated findings guide subsequent viewpoint selection, and acceptance requires both types of checks to satisfy their respective criteria. If verification concludes without acceptance, the acquired observations and verification findings are retained to guide subsequent exploration and answer revision.

Our main contributions are summarized as follows:

- We identify a fundamental failure mode in embodied understanding, where an interpretation can appear plausible before it is sufficiently supported by environmental evidence, causing agents to commit before their understanding has been adequately tested against the environment.
- We propose a verification-based interaction policy that shifts post-candidate interaction from answer seeking to hypothesis testing, and instantiate it in the training-free Sure-EQA framework.
- Experiments on A-EQA (Majumdar et al., 2024) and EXPRESS-Bench (Jiang et al., 2025) demonstrate consistent improvements in answer quality and path-weighted performance, while substantially reducing missing-evidence cases.

2 Related Work

Active Exploration for Embodied Understanding. Active exploration supports embodied understanding by acquiring information required to resolve a task (Das et al., 2018a; Gordon et al., 2018). Since this information may be distributed across multiple targets and action sequences, observation acquisition must be coordinated over a trajectory (Das et al., 2018b; Yu et al., 2019). Semantic priors guide this process by identifying locations likely to contain useful information (Zhou et al., 2023; Yokoyama et al., 2024). Selecting among these locations also requires considering their spatial context and relationships to the navigation goal (Wu et al., 2024; Huang et al., 2026). Scene representations and memory maintain this context across observations, allowing subsequent reasoning and navigation to draw on accumulated experience (Saxena et al., 2024; Yang et al., 2025b; Anwar et al., 2025). Recent predictive planning methods extend this process by anticipating informative observations and revising exploration decisions (Yuan et al., 2026; Zhou et al., 2026). These advances primarily organize interaction around acquiring and refining task-relevant environmental understanding. However, they do not explicitly require the agent to scrutinize an interpretation it already considers sufficient for task completion. This limitation therefore motivates our verification-based policy that redirects subsequent interaction toward examining the evidential basis of the agent's existing interpretation.

Verification for Embodied Understanding. Verification provides a basis for an embodied agent to assess whether an its environmental interpretation requires revision. In static settings, verification and self-correction reassess generated responses using the available inputs (Madaan et al., 2023; Lee et al., 2024; Zhou et al., 2024). These methods check generated claims and reasoning against supplied images (Yin et al., 2024b; Sahu et al., 2024; Liao et al., 2025). Such checks can correct the interpretation of existing evidence, but cannot inspect scene content absent from the inputs. Interactive embodied settings relax this restriction through additional observations. A prominent line of work focuses on target identification and goal-directed navigation. By inspecting candidate targets from additional viewpoints (Lei et al., 2024; Yin et al., 2024a), these methods assess correspondence with the navigation goal and determine whether to confirm a target or resume search (Cao et al., 2025; Yin et al., 2025). An agent's environmental understanding, however, integrates judgments about objects, properties, and their relations within a scene. Verification centered on a navigation goal only assesses the judgments needed for target confirmation, while the embodied understanding may depend on functional or contextual judgments beyond the goal. We instead center active environmental verification on the interpretation expressed in the agent's proposed answer. This shifts verification from assessing whether a target satisfies a navigation goal to testing whether the agent's understanding is justified by environmental evidence.

3 Method

We instantiate the proposed verification-based policy in Sure-EQA using EQA as a testbed. Sure-EQA addresses premature commitment by separating the answerer's decision to propose an answer from the decision to accept it. The candidate and its selected support make the agent's interpretation explicit, allowing further interaction to examine its evidential basis before commitment.

3.1 From Exploration to Verification

To study how verification changes exploration-based interaction, we consider a general EQA agent with an answerer and a planner sharing two memories. Visual memory $\mathcal{O}$ stores observations, while structural memory $\mathcal{H}$ records scene information and exploration history.

Answerer. Given a question q, the answerer assesses the evidence stored in both memories:

$$a \leftarrow \text{Answerer}(q, \mathcal{O}, \mathcal{H}). \tag{1}$$

It returns $a = (\hat{y}, S)$ when the answerer judges the evidence sufficient to answer q. Here, $\hat{y}$ is the candidate answer, and $S \subseteq \mathcal{O}$ contains its selected supporting images. Otherwise, it returns `Continue Exploration` to request further observations.

Planner. When further exploration is requested, we obtain target set $\mathcal{F}$ through truncated signed distance function (TSDF)-based frontier extraction (Ren et al., 2024). The planner uses the question

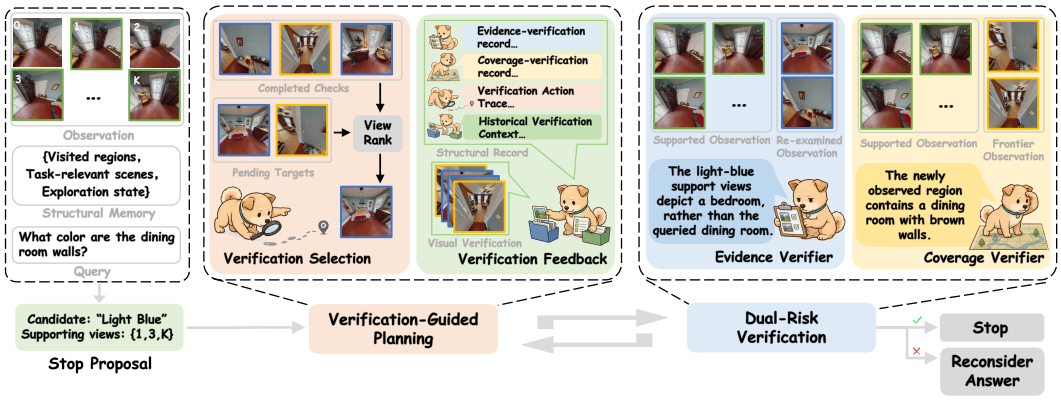


Figure 3: Overview of Sure-EQA. Following an exploration-based agent's stop proposal, verification findings guide further observation within a round, which ends with candidate acceptance or feedback for subsequent exploration and answer revision.

and both memories to select a target:

$$f \leftarrow \text{Planner}(q, \mathcal{O}, \mathcal{H}, \mathcal{F}), \qquad f \in \mathcal{F}. \tag{2}$$

The agent navigates toward f, acquires new observations, and updates both memories.

Under the exploration-based policy, returning $(\hat{y}, \mathcal{S})$ would terminate interaction. This decision relies on the answerer's own judgment of plausibility without further testing whether its interpretation is warranted, potentially leaving a gap between the available evidence and the evidence needed to establish the answer. Sure-EQA instead treats this output as a *stop proposal* that initiates a verification round. The interaction objective shifts from acquiring information to generate an answer to testing whether environmental evidence justifies the proposed candidate $\hat{y}$, so producing a plausible answer no longer suffices to terminate interaction. The candidate answer $\hat{y}$ and its selected support $\mathcal{S}$ remain fixed throughout the round. Subsequent observations are used to assess this interpretation rather than immediately incorporated into another candidate, providing a consistent object of verification as further evidence is acquired.

3.2 Dual-Risk Candidate Verification

Having formed a candidate, the agent needs to examine whether its understanding of the environment actually supports that answer. Simply forcing an exploration-based policy to continue exploring does not necessarily provide this verification. The agent may continue exploring and interpreting new observations based on its existing understanding, without examining whether the original evidence genuinely justifies the candidate. From a verification perspective, the existing interpretation should be the object of scrutiny rather than the premise for further inference. We therefore revisit how the available observations have been interpreted as support for the candidate and use additional observations to test whether this interpretation is justified and the evidence genuinely warrants the proposed answer.

We address two ways in which the interpretation may be mistaken. First, the supporting observations may have been misinterpreted. Occlusion, blur, and limited scene context can obscure details needed to establish the objects, attributes, and relations underlying the answer. Evidence verification (type E) therefore compares additional views with the original support to reassess these judgments. Second, even correctly interpreted support may omit relevant information outside the observed regions. Coverage verification (type C) examines newly observed regions for such information that could contradict or materially revise the candidate. These two complementary checks examine both the interpretation of available evidence and omissions in that evidence.

For each check i, let o_i denote the acquired new observation and $d_i \in \{E, C\}$ the selected target's type (see Section 3.3 for the target selection rule). The corresponding VLM verifier returns a judgment z_i and an explanation e_i:

$$(z_i, e_i) \leftarrow \text{Verifier}_{d_i}(q, \hat{y}, S, o_i, \mathcal{H}), \qquad z_i \in \{\texttt{SUPPORTED}, \texttt{CHANGED}, \texttt{UNCLEAR}\}. \tag{3}$$

SUPPORTED favors retaining the candidate, CHANGED calls for revision, and UNCLEAR indicates an inconclusive check. The explanation e_i states why the observation supports retaining the candidate, calls for revision, or leaves the check inconclusive.

The two verifier types are described below:

Evidence Verification. The evidence Verifier_E compares the new observation with S to check whether S has been correctly interpreted as support for $\hat{y}$. It establishes whether the views depict the same relevant object or region and reassesses the identity, attributes, states, and relationships underlying the candidate as additional details emerge. In Figure 3, the candidate "light blue" concerns dining-room walls, but the new view reveals that the supporting images depict a bedroom, exposing the mistaken association.

Coverage Verification. The coverage Verifier_C checks newly observed regions for key evidence omitted from S that contradicts the candidate or requires its revision. In the same example, a new observation reveals the previously unseen dining room with brown walls. This contradicts "light blue", calling for revision.

Verification itself relies on partial observations, so an individual judgment can also be inconclusive or misleading. To address this issue, we aggregate findings across checks into a quantitative answer-change risk score, allowing acceptance to be assessed from accumulated evidence rather than an isolated judgment. The score rises as evidence calling for revision accumulates and falls as observations support retaining the candidate. We maintain separate evidence and coverage risk scores because confirming the interpretation of existing support does not rule out answer-changing evidence in unseen regions.

For a verification target u of type $d \in \{E, C\}$, we define its answer-change risk as

$$r_d(u) = \frac{\alpha_0 + m_d^{\texttt{CHANGED}}(u)}{\alpha_0 + \beta_0 + m_d^{\texttt{CHANGED}}(u) + m_d^{\texttt{SUPPORTED}}(u)}. \tag{4}$$

Here, $m_d^{\texttt{SUPPORTED}}(u)$ measures accumulated support for retaining the candidate. The term $m_d^{\texttt{CHANGED}}(u)$ measures accumulated evidence calling for revision. The positive constants α_0 and β_0 set the initial risk to $\alpha_0/(\alpha_0 + \beta_0)$.

Let z denote either SUPPORTED or CHANGED. A finding's relevance varies across targets, while repeated observations from nearly identical poses can contribute redundant evidence. We therefore use pose similarity as a proxy for spatial relevance and weight each check by its geometric novelty relative to earlier checks of the same type and outcome. The accumulated evidence is

$$m_d^z(u) = \sum_{\substack{i:\, d_i = d \\ z_i = z}} k(u, u_i)\, w_i, \qquad z \in \{\texttt{SUPPORTED}, \texttt{CHANGED}\}. \tag{5}$$

Here, u_i denotes the actual camera position and viewing direction at which observation o_i was acquired. The geometric similarity $k(u, u_i)$ measures how closely this pose matches the target pose u. The weight w_i accounts for the geometric novelty of check i relative to earlier checks of the same type and outcome. An UNCLEAR outcome provides insufficient evidence for either retaining or revising the candidate. It therefore contributes to neither evidence sum.

3.3 Planning with Verification Findings

After each check, the updated answer-change risks and the availability of remaining verification targets determine how the interaction proceeds: whether to accept the candidate, continue verification, or return the accumulated findings to guide further exploration and answer revision.

Accepting the candidate. We accept the candidate when the risk at every verification target is at or below the threshold for its type:

$$r_d(u) \le \tau_d, \qquad \forall d \in \{E, C\}, \quad \forall u \in \mathcal{T}^d. \tag{6}$$

Here, $\mathcal{T}^d$ contains all pending and completed targets of type d in the current round. The thresholds τ_E and τ_C specify the acceptance criteria for evidence and coverage verification, respectively. If this

condition is not met and the verification round can continue, we use the updated risk scores to select the next verification target.

Continuing verification. Further observations should address unresolved concerns about the candidate while avoiding redundant observations. We therefore prioritize targets with high answer-change risk and geometric novelty, while discounting travel cost to use the interaction budget efficiently. Let $\mathcal{U}$ contain all reachable targets of both verification types that have not yet been checked in the current round. We select the next target as

$$u^\star = \arg\max_{u \in \mathcal{U}} \frac{r(u)\max\{\epsilon, \nu(u)\}}{1 + c(u)}. \tag{7}$$

Here, $r(u) = r_d(u)$ for a target of type $d \in \{E, C\}$. The novelty $\nu(u)$ measures how much the target's camera position and viewing direction differ from those of earlier checks of the same type. The travel cost $c(u)$ is the planar Euclidean distance from the agent's current position to the target. The constant $\epsilon > 0$ sets a positive lower bound on the novelty term. The resulting verification findings update the risk scores for the next decision.

Verification Feedback. If either verification type exhausts its pending targets while any of its targets remains above threshold, the round ends. The agent then switches back to the exploration-based policy. If it continues seeking an answer while retaining the same mistaken interpretation, further exploration may produce another incorrect candidate. The agent therefore needs to understand why its interpretation conflicts with the available evidence or remains insufficiently supported. To provide this feedback, we retain all acquired observations $\mathcal{O}_{\text{ver}}$ and their verification record $\mathcal{H}_{\text{ver}} = \{(d_i, u_i, z_i, e_i) \mid o_i \in \mathcal{O}_{\text{ver}}\}$. Each record associates an observation with its verification type, viewpoint, judgment, and explanation, preserving the evidence and reasoning needed to reassess the previous answer. We incorporate the images into visual memory and the verification records into structural memory as

$$\mathcal{O}^+ \leftarrow \mathcal{O} \cup \mathcal{O}_{\text{ver}}, \qquad \mathcal{H}^+ \leftarrow \mathcal{H} \cup \{\mathcal{H}_{\text{ver}}\}. \tag{8}$$

Using the updated memories, the answerer can request further exploration or propose another candidate for a new verification round. The agent thus alternates between exploration and verification until a candidate passes verification or the shared interaction budget is exhausted. Further methodological details are provided in Appendix A.

4 Experiments

4.1 Experimental Setups

Benchmarks. We evaluate Sure-EQA on A-EQA, the active setting of OpenEQA (Majumdar et al., 2024), and EXPRESS-Bench (Jiang et al., 2025). Both require agents to answer questions using observations acquired through environmental interaction, allowing us to examine whether verification improves answers that agents are already prepared to return. Following prior evaluations, we use the 184-question A-EQA subset and the 409-question EXPRESS-Bench subset to maintain consistency with existing benchmark results.

Evaluation metrics. We evaluate three aspects of performance. Answer correctness is measured by LLM-Match on A-EQA and C^* on EXPRESS-Bench. Grounded answer quality is measured by C on EXPRESS-Bench, which jointly accounts for correctness and visual support. To assess answer quality relative to travel cost, we report LLM-SPL on A-EQA and E_{path} on EXPRESS-Bench, which weight their respective answer-quality scores by path efficiency. To account for the travel required by verification, we compute both path-weighted metrics over the full trajectory, including exploration and verification movements. We also report d_T, the mean geodesic distance in meters from the final agent position to the reference target.

Implementation. Experiments run in Habitat (Savva et al., 2019) with Habitat-Matterport 3D (HM3D) scenes (Ramakrishnan et al., 2021). Sure-EQA builds on the exploration and memory

Table 1: Comparative evaluation on A-EQA and EXPRESS-Bench. Bold and underlined values indicate the best and second-best results among embodied agents, respectively. † indicates a VLM fine-tuned on 3D data. Exp. stands for exploration method.

(a) A-EQA

Method	VLM	LLM-Match ↑	LLM-SPL ↑
Human Agent	–	85.1	N/A
Blind LLMs			
GPT-4	–	35.5	N/A
GPT-4o	–	35.9	N/A
Proprietary VLMs			
Memory-EQA (Zhai et al., 2025a)	GPT-4o	36.8	–
ConceptGraphs (Gu et al., 2024)	GPT-4o	47.2	33.3
Explore-EQA (Ren et al., 2024)	GPT-4o	46.9	23.4
Fine-EQA (Jiang et al., 2025)	GPT-4o-mini	43.3	29.2
MTU3D (Zhu et al., 2025)	GPT-4o	51.1	42.6
3D-Mem (Yang et al., 2025b)	GPT-4o	52.6	42.0
Open-Source VLMs			
Memory-EQA (Zhai et al., 2025a)	Qwen2-VL 7B	27.4	–
ToolEQA† (Zhai et al., 2025b)	Qwen2.5-VL 7B	35.5	–
Pred-EQA (Yuan et al., 2026)	Qwen3-VL 8B	53.3	<u>48.5</u>
Sure-EQA	Qwen2.5-VL 7B	<u>53.9</u>	45.2
Sure-EQA	Qwen3-VL 8B	**58.7**	**52.8**

(b) EXPRESS-Bench

Method	VLM	$C\uparrow$	$C^*\uparrow$	$E_{\text{path}}\uparrow$	$d_T\downarrow$
Human Agent	–	N/A	83.99	N/A	N/A
Blind LLMs					
DeepSeek-V3	–	N/A	59.15	N/A	N/A
GPT-4	–	N/A	58.96	N/A	N/A
LLaMA-3-8B	–	N/A	57.25	N/A	N/A
Proprietary VLMs					
Random Exp. (Jiang et al., 2025)	GPT-4o-mini	36.95	62.75	12.06	7.26
Frontier Exp. (Jiang et al., 2025)	GPT-4o-mini	38.60	60.24	14.55	6.64
Goal Exp. (Jiang et al., 2025)	GPT-4o-mini	38.54	63.34	12.74	6.46
Fine-EQA (Jiang et al., 2025)	Qwen-VL-Plus	40.02	63.06	17.57	6.41
Fine-EQA (Jiang et al., 2025)	GPT-4o-mini	40.55	63.95	16.22	6.43
Open-Source VLMs					
Fine-EQA (Jiang et al., 2025)	Janus-Pro 7B	39.88	63.35	20.86	6.18
ToolEQA† (Zhai et al., 2025b)	Qwen2.5-VL 7B	42.21	65.77	25.82	**5.25**
Pred-EQA (Yuan et al., 2026)	Qwen3-VL 8B	<u>52.58</u>	70.54	<u>47.66</u>	<u>5.64</u>
Sure-EQA	Qwen2.5-VL 7B	51.71	<u>72.27</u>	41.49	6.57
Sure-EQA	Qwen3-VL 8B	**57.31**	**78.19**	**50.69**	6.14

modules of Pred-EQA (Yuan et al., 2026) without additional model training. Our main configuration uses Qwen3-VL 8B (Bai et al., 2025), with additional results using Qwen2.5-VL 7B. The answerer and both verifiers share the same backbone and use role-specific prompts. For the main evaluation, exploration and verification share a 50-step interaction budget, and verification parameters remain the same across both benchmarks. Further experimental details are provided in Appendix A.

4.2 Comparative Evaluation

Table 1 shows that Sure-EQA with Qwen3-VL 8B achieves the highest answer-quality and path-weighted scores among the compared automated agents. Relative to Pred-EQA with the same backbone, it improves LLM-Match by 5.4 points on A-EQA and C^* by 7.65 points on EXPRESS-Bench. They support improving embodied answering not only by seeking information from which an answer can be formed, but also by examining whether a proposed answer is sufficiently supported before committing to it. Importantly, LLM-SPL and E_{path} account for the full trajectory, including verification movements, indicating that the answer-quality benefits remain after verification travel is charged, rather than arising from additional interaction whose cost is excluded. The improvement in C further shows that the gains persist when correctness is evaluated jointly with visual support.

4.3 Ablation Study

Dual-risk verification. Table 2 evaluates the benefits of checking existing evidence and potentially missing information. Both branches improve over the unverified baseline, while their combination performs best across all metrics, gaining 2.7 points in LLM-Match and 3.5 in C^* over the better single branch. This pattern is consistent with their complementary roles: re-examining existing evidence can correct misinterpretations, whereas exploring unseen regions can reveal missing information. The results support checking both risks before accepting a candidate.

Verification selection. Table 3 shows how viewpoint selection affects verification. Removing risk from viewpoint selection reduces LLM-Match by 4.8 points with the acceptance rule unchanged. Omitting viewpoint novelty also lowers performance, supporting the value of perspectives that differ from previous checks. Removing navigation cost largely preserves LLM-Match but reduces LLM-SPL by 4.1 points. The risk ablation holds acceptance fixed, so its decline supports using verification findings to guide observation, beyond their role in determining when to stop.

Verification feedback. The feedback ablations in Table 3 separately omit verification records from structural memory and new verification images from visual memory. Both retain verification using newly acquired observations but restrict the information available to subsequent reasoning. The declines distinguish performing a check from retaining its findings for subsequent reasoning. Images preserve the visual basis for revision, while linked records explain how the observations bear on

the previous answer. Feedback thus connects candidate evaluation to subsequent answer formation. Feedback is therefore effective even when the current round does not establish an acceptable answer. By retaining why a candidate was challenged or remained unresolved, subsequent reasoning can revisit the assumptions behind it instead of merely generating another answer from an expanded image collection.

Different candidate generators. Finally, we evaluate whether the gains depend on the system that produces the initial candidate. Table 4 evaluates Memory-EQA, ToolEQA, Fine-EQA, and Pred-EQA with the same Qwen3-VL 8B backbone. Sure-EQA improves both LLM-Match and LLM-SPL for all four, with mean gains of 8.0 and 5.4 points, respectively. These gains suggest that verification complements different exploration and memory designs rather than requiring a particular candidate generator.

4.4 Verification Analysis

Beyond continued exploration. Figure 4(a) compares strategies starting from the same baseline candidate-answer states on A-EQA. Baseline returns the candidate, while Self-check reassesses it using existing observations and one additional VLM call. Continued exploration prompts the original planner to seek counterevidence to the candidate, whereas Sure-EQA organizes verification through complementary evidence and coverage checks with risk-based control. For these two policies, additional observation steps are matched per question, and the same answerer uses the updated memories. Sure-EQA exceeds continued exploration by 4.8 LLM-Match points, while Self-check offers only a small improvement over Baseline. With the same initial candidates and matched observation steps, this improvement cannot be explained solely by acquiring more observations. The improvement supports structured verification as a more effective way to reassess proposed answers under matched observation budgets.

Table 2: Effectiveness of dual-risk verification.

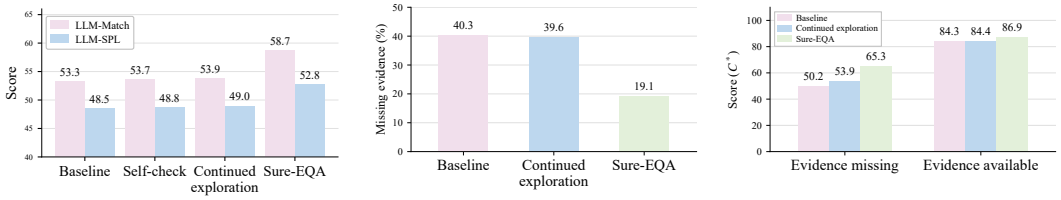
Evi.	Cov.	A-EQA		EXPRESS-Bench	
		LLM-Match↑	LLM-SPL↑	C^* ↑	E_{path} ↑
–	–	53.3	48.5	70.5	47.7
✓	–	54.6	49.6	73.9	49.0
–	✓	56.0	50.9	74.7	48.3
✓	✓	**58.7**	**52.8**	**78.2**	**50.7**

Table 3: Ablations of verification selection and feedback.

Variant	A-EQA		EXPRESS-Bench	
	LLM-Match↑	LLM-SPL↑	C^* ↑	E_{path} ↑
Baseline (Qwen3-VL 8B)	53.3	48.5	70.5	47.7
Verification Selection				
w/o. View Novelty	55.0	50.1	75.9	49.1
w/o. Navigation Cost	58.2	48.7	76.3	46.8
w/o. Risk Score	53.9	49.1	73.5	48.1
Verification Feedback				
w/o. Structural Record	54.3	49.5	74.0	48.3
w/o. Visual Feedback	55.7	50.1	75.4	48.9
Sure-EQA	**58.7**	**52.8**	**78.2**	**50.7**

Table 4: Effectiveness of Sure-EQA with different candidate generators on A-EQA. * indicates our reimplementation with the same backbone.

Method (Qwen3-VL 8B)	Sure-EQA	LLM-Match↑	LLM-SPL↑
Memory-EQA* (Zhai et al., 2025a)	–	30.6	27.4
ToolEQA* (Zhai et al., 2025b)	–	38.7	29.8
Fine-EQA* (Jiang et al., 2025)	–	40.1	28.3
Pred-EQA (Yuan et al., 2026)	–	53.3	48.5
Memory-EQA (Zhai et al., 2025a)	✓	39.1	32.7
ToolEQA (Zhai et al., 2025b)	✓	46.3	35.2
Fine-EQA (Jiang et al., 2025)	✓	50.5	34.9
Pred-EQA (Yuan et al., 2026)	✓	**58.7**	**52.8**



(a) Answer quality under different interaction strategies.

(b) Missing visual evidence at final answering.

(c) Answer quality by initial evidence availability.

Figure 4: Verification analysis on A-EQA (a) and EXPRESS-Bench (b,c) with Qwen3-VL 8B. Evidence availability is determined by annotated question-relevant object visibility in the answerer's observations. Panel (c) reports mean C^* for fixed groups with evidence missing ($n = 165$) or available ($n = 244$) at Baseline's initial stop proposal.

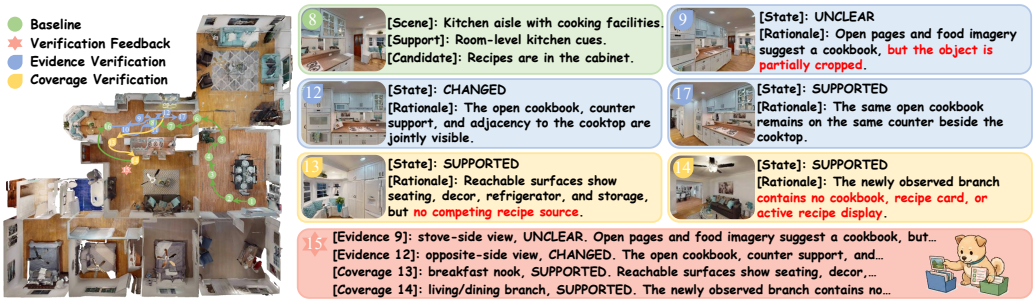


Figure 5: Qualitative verification trace for the query "Where can I get recipes for what to cook?" Sure-EQA replaces the baseline's wrong candidate with the visually grounded answer "The recipes are on the kitchen counter."

Evidence acquisition analysis. On EXPRESS-Bench, we assess evidence availability through the visibility of annotated question-relevant objects. Sure-EQA reduces missing evidence from 40.3% to 19.1%, versus 39.6% for continued exploration (Figure 4(b)). For fixed groups defined at Baseline's initial stop proposal, its C^* gains over Baseline are 15.1 points when evidence is initially missing and 2.6 when available (Figure 4(c)). The joint improvement in evidence availability and answer quality supports verification as an intervention in evidence acquisition, rather than merely a reformulation of the answer.

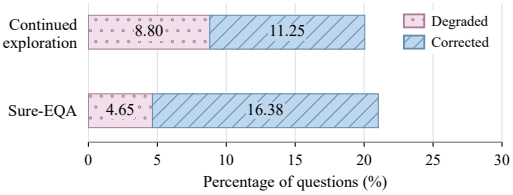


Figure 6: Answer-correctness transitions on EXPRESS-Bench ($N = 409$). Corrected and degraded denote incorrect-to-correct and correct-to-incorrect transitions, respectively.

Correcting mistaken interpretations. We analyze answer-correctness transitions on EXPRESS-Bench from the same baseline stop proposals, comparing Sure-EQA with continued exploration under matched additional observation budgets. Figure 6 shows that Sure-EQA corrects more initially mistaken answers while preserving more initially correct ones. This yields a net increase in the fully correct answer rate of 11.74 percentage points, compared with 2.44 for continued exploration. The results support the shift from answer seeking to hypothesis testing in exposing and revising plausible but mistaken interpretations that agents were already prepared to accept, thereby improving embodied understanding before commitment.

Tracing answer revision. Figure 5 traces how verification revises an unsupported cabinet prediction. A cropped view yields `UNCLEAR`, while another viewpoint reveals a cookbook on the counter and yields `CHANGED`. After coverage checks find no competing recipe source, the observations and rationales are retained at round completion to guide candidate regeneration. A later view yields `SUPPORTED`, illustrating how environmental evidence challenges a plausible but mistaken interpretation and how retained findings guide its revision.

5 Conclusion

We propose a verification-based policy that shifts an embodied agent's interaction objective from answer seeking to hypothesis testing once a candidate answer is formed. Using EQA as a testbed, we introduce the training-free Sure-EQA framework to test the proposed interpretation through complementary checks of existing support and potentially missing evidence. Accumulated findings guide further verification and candidate acceptance, while retained feedback informs subsequent exploration and answer revision. Our experimental results show that directing further interaction toward testing a proposed interpretation helps agents acquire missing evidence and revise mistaken answers. These findings support organizing post-candidate interaction through complementary checks and accumulated verification findings to improve embodied understanding beyond candidate-conditioned continued exploration.

AI USE STATEMENT

Codex was used to assist with manuscript drafting and language refinement, discussions of experimental design, preparation of plotting code, and creation of schematic illustrations. All AI-assisted text, code, and visual materials were reviewed and revised as needed, and we determined which suggestions and materials to incorporate. We confirm that all research decisions and conclusions remained under our control and take full responsibility for the final content of this work, including its claims, experimental results and code.

REFERENCES

Abrar Anwar, John Welsh, Joydeep Biswas, Soha Pouya, and Yan Chang. Remembr: Building and reasoning over long-horizon spatio-temporal memory for robot navigation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2838–2845. IEEE, 2025.

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

Yihan Cao, Jiazhao Zhang, Zhinan Yu, Shuzhen Liu, Zheng Qin, Qin Zou, Bo Du, and Kai Xu. Cognav: Cognitive process modeling for object goal navigation with llms. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9550–9560. IEEE, 2025.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2135–213509. IEEE, 2018a.

Abhishek Das, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Neural modular control for embodied question answering. In *Conference on robot learning*, pp. 53–62. PMLR, 2018b.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 91–104. IEEE, 2025.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pp. 226–231, 1996.

Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. Iqa: Visual question answering in interactive environments. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4089–4098. IEEE, 2018.

Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5021–5028. IEEE, 2024.

Xun Huang, Shijia Zhao, Yunxiang Wang, Xin Lu, Wanfa Zhang, Rongsheng Qu, Weixin Li, Yunhong Wang, and Chenglu Wen. Msgnav: Unleashing the power of multi-modal 3d scene graph for zero-shot embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37154–37163, 2026.

Kaixuan Jiang, Yang Liu, Weixing Chen, Jingzhou Luo, Ziliang Chen, Ling Pan, Guanbin Li, and Liang Lin. Beyond the destination: A novel benchmark for exploration-aware embodied question answering. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9091–9101. IEEE, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. Volcano: mitigating multimodal hallucination through self-feedback guided revision. 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. 391–404, 2024.

Xiaohan Lei, Min Wang, Wengang Zhou, Li Li, and Houqiang Li. Instance-aware exploration-verification-exploitation for instance imagegoal navigation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16329–16339. IEEE, 2024.

Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, and David Acuna. Can large vision-language models correct semantic grounding errors by themselves? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14667–14678. IEEE, 2025.

Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. Probing visual language priors in vlms. *arXiv preprint arXiv:2501.00569*, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16488–16498. IEEE, 2024.

Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021.

Allen Z Ren, Jaden Clark, Anushri Dixit, Masha Itkina, Anirudha Majumdar, and Dorsa Sadigh. Explore until confident: Efficient exploration for embodied question answering. *arXiv preprint arXiv:2403.15941*, 2024.

Pritish Sahu, Karan Sikka, and Ajay Divakaran. Pelican: Correcting hallucination in vision-llms via claim decomposition and program of thought verification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8228–8248, 2024.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9338–9346. IEEE, 2019.

Saumya Saxena, Blake Buchanan, Chris Paxton, Peiqi Liu, Bingqing Chen, Narunas Vaskevicius, Luigi Palmieri, Jonathan Francis, and Oliver Kroemer. Grapheqa: Using 3d semantic scene graphs for real-time embodied question answering. *arXiv preprint arXiv:2412.14480*, 2024.

Andrew Szot, Bogdan Mazoure, Omar Attia, Aleksei Timofeev, Harsh Agrawal, Devon Hjelm, Zhe Gan, Zsolt Kira, and Alexander Toshev. From multimodal llms to generalist embodied agents: Methods and lessons. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10644–10655. IEEE, 2025.

Pengying Wu, Yao Mu, Bingxian Wu, Yi Hou, Ji Ma, Shanghang Zhang, and Chang Liu. Voronav: Voronoi-based zero-shot object navigation with large language model. *arXiv preprint arXiv:2401.02695*, 2024.

Tao Wu, Chuhao Zhou, Guangyu Zhao, Haozhi Cao, Yewen Pu, and Jianfei Yang. When robots should say"i don't know": Benchmarking abstention in embodied question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15266–15275, 2026.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2087–2098. IEEE, 2025.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*, 2025a.

Yuncong Yang, Han Yang, Jiachen Zhou, Peihao Chen, Hongxin Zhang, Yilun Du, and Chuang Gan. 3d-mem: 3d scene memory for embodied exploration and reasoning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17294–17303. IEEE, 2025b.

Hang Yin, Xiuwei Xu, Zhenyu Wu, Jie Zhou, and Jiwen Lu. Sg-nav: Online 3d scene graph prompting for llm-based zero-shot object navigation. *Advances in neural information processing systems*, 37:5285–5307, 2024a.

Hang Yin, Xiuwei Xu, Linqing Zhao, Ziwei Wang, Jie Zhou, and Jiwen Lu. Unigoal: Towards universal zero-shot goal-oriented navigation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19057–19066. IEEE, 2025.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024b.

Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. Vlfm: Vision-language frontier maps for zero-shot semantic navigation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 42–48. IEEE, 2024.

Licheng Yu, Xinlei Chen, Georgia Gkioxari, Mohit Bansal, Tamara L Berg, and Dhruv Batra. Multi-target embodied question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6302–6311. IEEE, 2019.

Bowen Yuan, Sisi You, and Bing-Kun Bao. Predict before you explore: Predictive planning with specialized memory for embodied question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29610–29619, 2026.

Mingliang Zhai, Zhi Gao, Yuwei Wu, and Yunde Jia. Memory-centric embodied question answering. *arXiv preprint arXiv:2505.13948*, 2025a.

Mingliang Zhai, Hansheng Liang, Xiaomeng Fan, Zhi Gao, Chuanhao Li, Che Sun, Xu Bin, Yuwei Wu, and Yunde Jia. Multi-step reasoning for embodied question answering via tool augmentation. *arXiv preprint arXiv:2510.20310*, 2025b.

Kaiwen Zhou, Kaizhi Zheng, Connor Pryor, Yilin Shen, Hongxia Jin, Lise Getoor, and Xin Eric Wang. Esc: Exploration with soft commonsense constraints for zero-shot object navigation. In *International conference on machine learning*, pp. 42829–42842. PMLR, 2023.

Weijie Zhou, Xuantang Xiong, Yi Peng, Manli Tao, Chaoyang Zhao, Honghui Dong, Ming Tang, and Jinqiao Wang. Physvlm-avr: Active visual reasoning for multimodal large language models in physical environments. *Advances in Neural Information Processing Systems*, 38:3873–3906, 2026.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations*, volume 2024, pp. 56969–56998, 2024.

Ziyu Zhu, Xilin Wang, Yixuan Li, Zhuofan Zhang, Xiaojian Ma, Yixin Chen, Baoxiong Jia, Wei Liang, Qian Yu, Zhidong Deng, et al. Move to understand a 3d scene: Bridging visual grounding and exploration for efficient and versatile embodied navigation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8120–8132. IEEE, 2025.

A APPENDIX

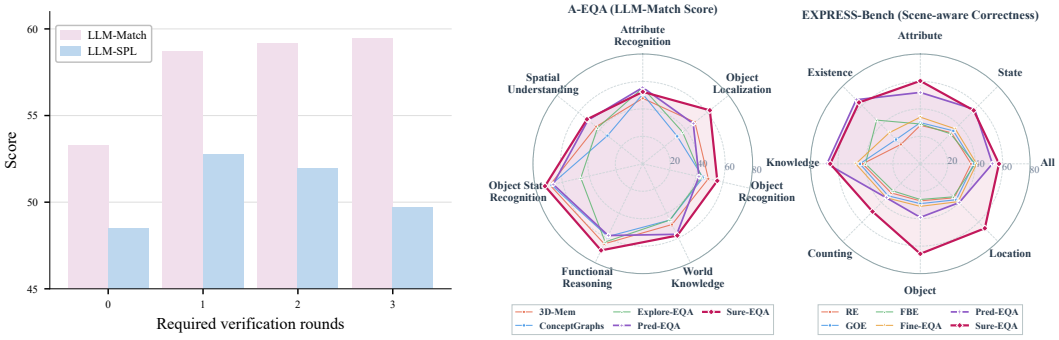
A.1 EXTENDED EXPERIMENTS

Pilot evidence intervention. We conduct the pilot study on the full EXPRESS-Bench dataset, comprising 2,044 question–trajectory pairs, using EQA agents with Qwen2.5-VL 7B, Qwen3-VL 8B, and Qwen3-VL 32B backbones. For each backbone, we manually assess whether the annotated question-relevant objects are visible in the observations used to propose stopping, using the annotated object locations, agent poses, and depth information. This visibility assessment defines evidence availability in both the pilot study and the main-text evidence analysis. Cases with missing evidence account for 51.3%, 44.2%, and 38.6% of the evaluated samples, respectively.

To obtain the missing evidence, we use the benchmark-provided locations of question-relevant objects and manually adjust the agent's position and viewing direction to capture key evidence frames depicting those objects. We insert these frames into the available observations for each backbone's corresponding evidence-missing subset and compare answer correctness before and after insertion. The fully correct answer rates increase from 27.7% to 60.2%, from 30.4% to 72.9%, and from 35.8% to 80.5%, respectively. These paired comparisons show that supplying missing visual evidence can substantially improve answers that agents were otherwise ready to accept. The annotated object locations and manual viewpoint control are used for this diagnostic intervention. Sure-EQA acquires additional observations through its own exploration and verification process.

The continued-exploration control in Figure 1 follows the same protocol as in Figure 4. Starting from the same baseline candidate-answer state, the original exploration planner receives the candidate and is explicitly instructed to seek observations that could refute it. It selects subsequent viewpoints using its exploration policy, with additional observation steps matched to Sure-EQA for each question. The same answerer then generates an updated answer using the resulting memories. This comparison therefore evaluates the benefit of Sure-EQA's verification mechanism beyond explicitly directing the original planner to seek counterevidence. For each backbone, this control is evaluated on the same evidence-missing subset used for manual evidence insertion.

Required verification rounds. We vary the number of successful verification rounds required before a verified stop is permitted. Each round passes when both risks are at or below their corresponding threshold. Setting i permits termination after i successful rounds, whereas $i = 0$ denotes the baseline without verification. As shown in Figure 7a, one successful verification round produces the main gain in LLM-Match over returning the unverified baseline answer. Requiring additional rounds yields only marginal improvements in LLM-Match while reducing LLM-SPL. These results support one successful verification round as an effective default for improving answer quality while limiting the travel required for additional evidence acquisition.



(a) Effect of the number of required verification rounds on A-EQA.

(b) Category-wise performance in LLM-Match on A-EQA and scene-aware correctness C on EXPRESS-Bench.

Figure 7: Verification analysis at different granularities. (a) Performance under different required verification rounds. (b) Category-wise comparison on A-EQA and EXPRESS-Bench.

Category-wise performance. Figure 7b decomposes the improvements by question category. On A-EQA, Sure-EQA outperforms the compared methods across nearly all categories, particularly in Object Localization, Object Recognition and Functional Reasoning. On EXPRESS-Bench, the

clearest gains occur for Location, Object, and Counting. These categories require precise entity identification, while partial observations can leave key evidence missing or ambiguous. The gains are consistent with the role of evidence verification in re-examining ambiguous cues from alternative viewpoints and coverage verification in acquiring potentially missing information. The results support testing the interpretation underlying an answer across multiple aspects of environmental understanding.

Adaptive verification budgets. We assess whether risk-based stopping allocates verification steps more effectively than a fixed budget. On all 184 A-EQA questions, we compare Sure-EQA with fixed-budget variants starting from the same baseline candidate-answer states. Both retain the same rules for viewpoint selection, verifier evaluation, memory feedback, candidate revision, and transitions between exploration and verification. They differ in how verification effort is allocated before termination. Sure-EQA uses accumulated verification risk to trigger verified stopping, whereas the fixed-budget variants require exactly K additional verification steps before final answering. Feedback and candidate revision follow the same round-level procedure throughout verification in both settings. We vary $K \in \{1, 3, 5, 7, 9, 11\}$. This diagnostic allows the fixed-budget variants to complete these additional steps beyond the standard 50-step episode limit, while Sure-EQA retains its standard episode budget.

As shown in Figure 8, fixed-budget verification achieves 55.8 LLM-Match at $K = 3$ and 58.6 at $K = 9$. Sure-EQA achieves 58.7 with an average of 3.36 verification steps over all 184 questions, including those with zero verification steps. It therefore uses 62.7% fewer verification steps than $K = 9$ at similar answer quality. Increasing the fixed budget from 9 to 11 steps yields only a further 0.4-point gain. These results support risk-based stopping as a means of allocating verification effort adaptively, achieving comparable answer quality with fewer checks while retaining the same feedback and candidate-revision procedure.

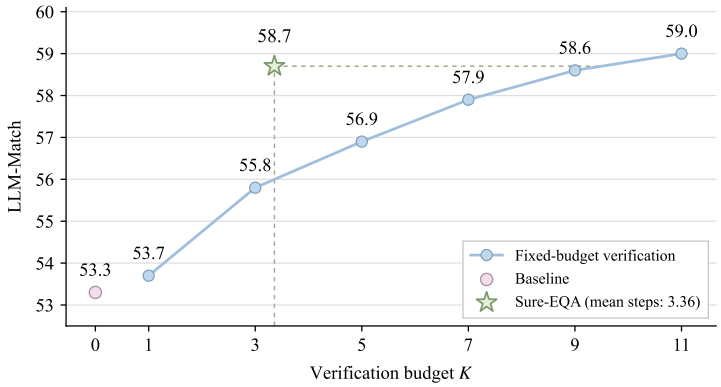


Figure 8: Answer quality under adaptive and fixed verification budgets on A-EQA with Qwen3-VL 8B. Blue markers show fixed-budget verification, where every question receives exactly K additional verification steps. The green star shows Sure-EQA at its mean verification length of 3.36 steps and LLM-Match of 58.7. Baseline is shown at $K = 0$. The mean is computed over all 184 questions, including those with zero verification steps.

Verification control. We examine whether the risk-based control rule improves the correctness of accepted answers while limiting verification cost. We evaluate this on all 184 A-EQA questions using Qwen3-VL 8B. Specifically, we compare Sure-EQA with a Count-based alternative whose branch score is $\widetilde{R}_d = (2 + N_d^{\text{CHANGED}})/(3 + N_d^{\text{CHANGED}} + N_d^{\text{SUPPORTED}})$, where the two counts record completed checks of type d in the current round. Both methods retain the same risk-based viewpoint-selection scoring, verifier prompts, memory feedback, and 50-step episode budget. Each method uses its respective branch scores to determine candidate acceptance and whether an exhausted branch ends the current round with feedback. We evaluate shared branch thresholds $\tau_E = \tau_C = \tau \in \{0.45, 0.49, 0.53, 0.57, 0.61\}$.

Figure 10(a) compares answer quality with mean verification steps over all questions. At the default threshold $\tau = 0.53$, Sure-EQA achieves 58.7 LLM-Match with 3.36 verification steps. Count-

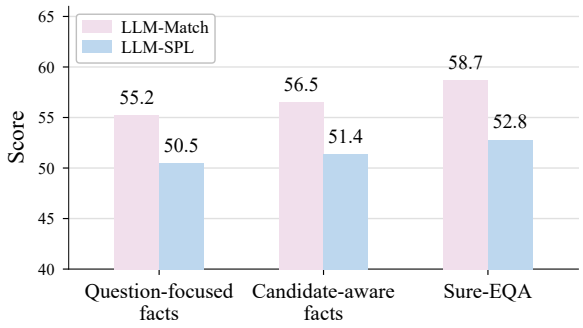


Figure 9: Comparison of feedback formulations on A-EQA. Question-focused facts and Candidate-aware facts summarize relevant observations without and with the initial candidate as explicit input, respectively. Sure-EQA explicitly evaluates the candidate against those observations. All variants share the same observations and trajectories.

based at $\tau = 0.45$ or 0.49 achieves 56.7 with 3.93 steps. Sure-EQA thus improves LLM-Match by 2.0 points while using 14.5% fewer verification steps. These settings also yield similar verified acceptance rates, with 156 questions accepted by Sure-EQA and 154 by Count-based. As shown in Figure 10(b), the proportions of accepted answers that are not fully correct are 41.67% and 46.75%, respectively. Sure-EQA therefore reduces this proportion by 5.1 percentage points while accepting a similar number of candidates.

Across the evaluated Sure-EQA settings, stricter thresholds reduce the proportion of incorrect accepted answers at the cost of more verification steps and fewer verified acceptances. The highest LLM-Match occurs at $\tau = 0.49$. The default $\tau = 0.53$ retains 58.7 LLM-Match with 23.3% fewer verification steps than this setting (3.36 versus 4.38). Among the five evaluated Sure-EQA thresholds, the default setting $\tau = 0.53$ also achieves the highest LLM-SPL of 52.8, supporting its balance between answer quality and navigation efficiency. At comparable acceptance rates, the proposed control rule reduces the proportion of accepted answers that are not fully correct relative to Count-based control. These results support risk-based control as a means of balancing acceptance quality and verification effort.

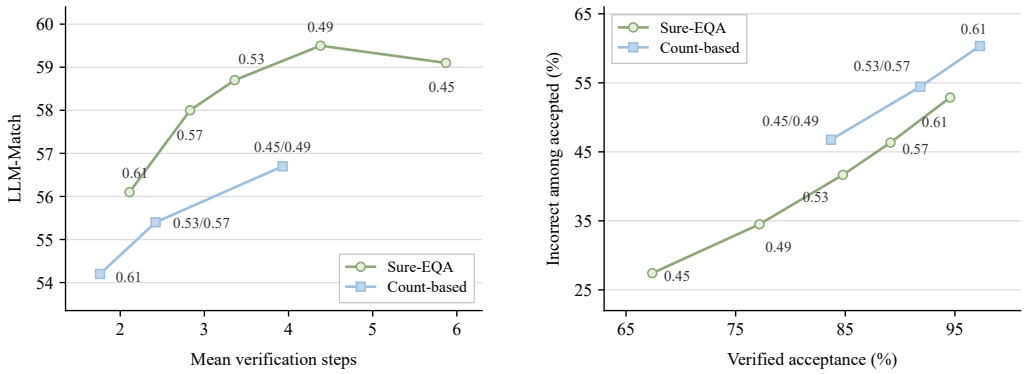
Beyond factual feedback. We study whether testing a proposed answer improves revision beyond factual descriptions of the same observations. On A-EQA, we compare Sure-EQA with two factual-feedback variants that share the same initial candidate answers, acquired observations, backbone, and final answerer. *Question-focused facts* summarizes question-relevant visual content without being given the candidate. *Candidate-aware facts* additionally receives the candidate but describes relevant facts without explicitly judging its correctness. Sure-EQA instead evaluates the observations against the proposed interpretation and retains the resulting verification judgments and rationales. Figure 9 shows that Sure-EQA outperforms both factual variants, improving LLM-Match by 2.2 points and LLM-SPL by 1.4 points over Candidate-aware facts. These results support explicitly reassessing a proposed interpretation against new observations as a more effective basis for answer revision than recording relevant facts alone.

Accumulation before feedback. We examine whether accumulating verification findings before feedback improves candidate revision. On A-EQA, we compare Sure-EQA with an immediate-feedback variant that retains the same risk-based acceptance criterion but triggers feedback and candidate revision as soon as a check returns `CHANGED`. Sure-EQA instead retains the candidate within each verification round and accumulates findings to determine whether to accept it or return feedback for revision. Both variants are evaluated on the same questions. We report LLM-Match, mean total interaction steps over complete episodes, and answer degradation. Interaction steps include both exploration and verification. Answer degradation is the percentage of all evaluated questions for which an initially fully correct answer becomes less than fully correct.

Table 5 shows that Sure-EQA improves LLM-Match by 4.5 points over immediate feedback while reducing mean total interaction steps by 43.1%. It also lowers answer degradation from 11.96% to 5.43%, indicating better preservation of initially correct answers. Verification itself relies on

Table 5: Comparison with immediate feedback on A-EQA. Answer degradation measures correct-to-incorrect transitions as a percentage of all evaluated questions.

Method	LLM-Match↑	Mean total interaction steps↓	Answer degradation (%)↓
Immediate feedback	54.2	19.81	11.96
Sure-EQA	**58.7**	**11.27**	**5.43**



(a) Answer quality versus verification cost.

(b) Incorrect answers versus verified acceptance.

Figure 10: Verification control and threshold sensitivity on A-EQA with Qwen3-VL 8B. (a) LLM-Match versus mean verification steps over all 184 questions. (b) The percentage of accepted answers that are not fully correct versus the percentage of all questions receiving verified acceptance. Point labels indicate the shared branch threshold $\tau_E = \tau_C = \tau$. Overlapping Count-based settings share a marker and display both thresholds. Lines connect evaluated settings.

partial observations, so an individual judgment may be misleading when relevant visual context is missing. Accumulating findings across checks before feedback reduces reliance on an isolated `CHANGED` judgment when guiding candidate revision. The lower degradation rate suggests that this accumulation helps mitigate the effects of partial-view ambiguity, improving final answer quality while limiting answer degradation and interaction costs.

Answer quality by termination outcome. Figure 11 groups the 184 A-EQA questions by Sure-EQA's termination outcome and compares Baseline and Sure-EQA on the same questions within each group. For 8 questions (4.3%), the budget is exhausted before verification begins, and mean LLM-Match remains 40.6. For 20 questions (10.9%), the budget is exhausted during verification. The acquired observations and verification records are fed back to generate the final answer, and mean LLM-Match increases from 51.3 to 52.5. The remaining 156 questions (84.8%) satisfy the joint verification condition, with mean LLM-Match increasing from 54.2 to 60.4. The improvement on questions that exhaust the budget during verification suggests that even partial checks can support revision of the current interpretation. By retaining the acquired observations and verification records, Sure-EQA helps the agent reconsider an answer it was initially ready to accept and ground its understanding in further environmental evidence.

Answer correctness at verified stopping. Within the subsets that reach verified stopping, we compare Baseline and Sure-EQA on 156 A-EQA questions and 380 EXPRESS-Bench questions. We use a strict binary criterion in which only fully correct answers are labeled correct. Partially correct answers are included in the incorrect group. Figure 12 shows that incorrect-to-correct transitions account for 18.59% and 16.58% of the respective subsets, whereas correct-to-incorrect transitions account for 5.13% and 4.21%. The fully correct answer rate increases from 44.87% to 58.33% on A-EQA and from 64.74% to 77.11% on EXPRESS-Bench. These paired transitions show that verification and subsequent revision correct more answers than they degrade within the verified-stop subsets. Nevertheless, 41.67% and 22.89% of accepted answers remain less than fully correct, in-

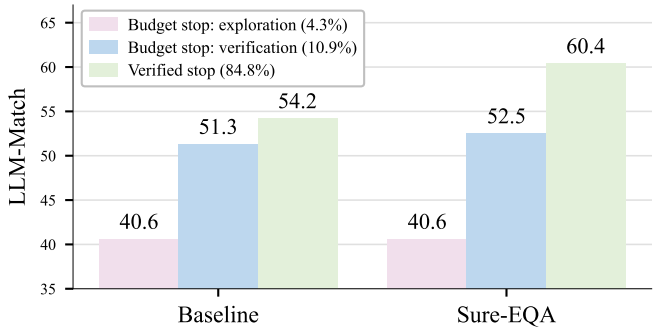


Figure 11: Answer quality by Sure-EQA's termination outcome on A-EQA. Both methods are evaluated on the same questions within each group. Bars report mean LLM-Match, and percentages indicate each group's share of the 184 evaluation questions.

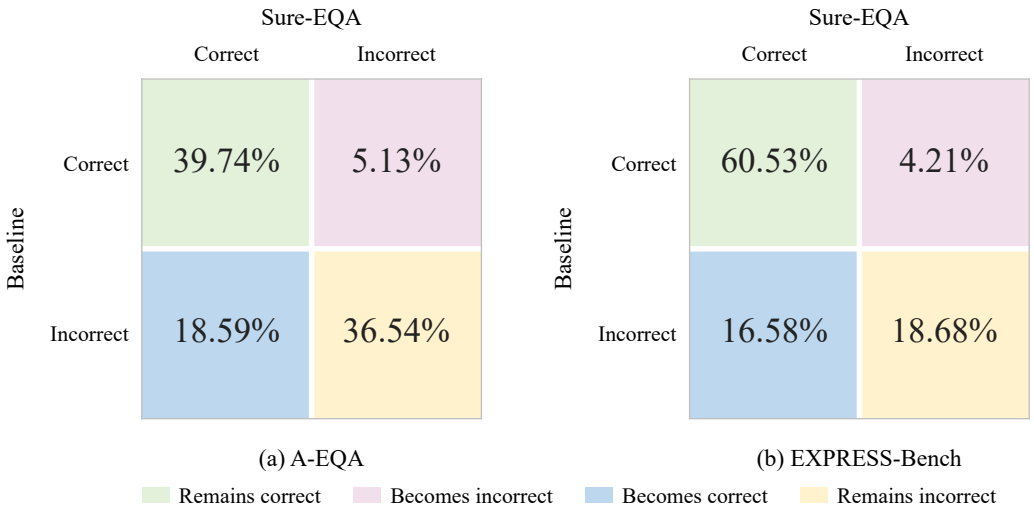


Figure 12: Answer-correctness transitions within the verified-stop subsets of (a) A-EQA ($N = 156$) and (b) EXPRESS-Bench ($N = 380$). Rows indicate baseline answer correctness, and columns indicate final Sure-EQA answer correctness. Each cell reports the percentage of the corresponding verified-stop subset. Correct denotes fully correct answers, and incorrect includes partially correct answers.

dicating that verified acceptance does not ensure complete correctness. Appendix A.2 illustrates remaining difficulties in preserving semantic specificity and obtaining accurate object counts.

A.2 Extended Qualitative Analysis

We complement the quantitative answer-correctness transition analysis with four qualitative cases in Figure 13. These cases compare baseline answers with Sure-EQA's final answers and illustrate all four transition types. Following the criterion used in the transition analysis, only fully correct answers are classified as correct; partially correct answers are included in the incorrect group.

Correct to correct. In Figure 13(a), the agent correctly answers that the ceiling lights in the living room are on. Verification produces conflicting judgments: a restricted doorway view leads one check to suggest that the lights are off at step 6, whereas step 7 and step 8 show illuminated ceiling fixtures. The candidate answer remains unchanged during these checks, and an additional view further supports retaining it. This case illustrates how further observation can help resolve disagreement among verification outcomes while preserving a correct interpretation, rather than treating an individual CHANGED outcome as an immediate answer revision.

Incorrect to correct. In Figure 13(b), the agent initially identifies two round decorative plates as the objects on top of the bedroom drawer. A new viewpoint at step 4 reveals a television and its stand on the drawer unit, contradicting the initial answer. This evidence is retained through refutation feedback and used in subsequent answering. Observation at step 7 shows the television, its stand, and the drawer together, supporting the corrected answer. The improvement comes from acquiring evidence that clarifies both the object's identity and its spatial relation to the drawer.

Correct to incorrect. In Figure 13(c), the baseline correctly identifies a closet with dark wooden louvered doors to the left of the staircase. Following verification and feedback, the final answer instead describes a door in that location. Although this description refers to a visible part of the scene, it omits the storage-closet identity required for a fully correct answer. The verification records support the door description without preserving the more specific interpretation. This case illustrates how VLM-based verification can accept a coarser description that preserves visible features but loses the semantic specificity needed to answer the question.

Incorrect to incorrect. In Figure 13(d), the agent initially reports that there are no paintings in the entryway at step 2. Coverage observations during steps 8–12 reveal previously missed artwork and challenge this absence claim. After feedback, the answer changes to four paintings and receives a `SUPPORTED` verification outcome, but the reference count is five. The additional observations lead the agent to revise the absence claim to a nonzero count, while the revised answer remains incorrect. This case highlights the difficulty of combining observations into an accurate count even after relevant scene content has been observed.

These cases illustrate how environmental verification can help retain or revise an agent's current interpretation. They also expose failures to preserve semantic specificity and to combine observations into an accurate count. Together with the transition analysis, they show the benefits and remaining limitations of using additional environmental evidence to reconsider candidate answers.

A.3 Verification Details

This appendix supplements the verification and planning procedures described in Sections 3.2 and 3.3. Appendix A.3.1 describes how the planning procedure constructs candidate viewpoints for evidence and coverage verification. Appendix A.3.2 specifies the geometric weighting used to aggregate verification findings and the novelty measure used to select subsequent targets. Appendix A.3.3 details target updates, candidate acceptance, round termination, and verification feedback.

A.3.1 Viewpoint Construction

The proposed interpretation is tested through observations acquired at reachable viewpoints. A verification target specifies a camera position and viewing direction. We construct evidence viewpoints around regions depicted in observations selected as support and coverage viewpoints at the boundaries of unexplored space. Here, t denotes the current interaction step. The support set $\mathcal{S}_t$ corresponds to $\mathcal{S}$ in the main text and remains fixed within a verification round.

Evidence viewpoint construction. Evidence targets are associated with the supporting observations in $\mathcal{S}_t$. For a support observation s, let $\mathbf{p}_s \in \mathbb{R}^2$ and $\mathbf{g}_s$ denote its horizontal camera position and unit viewing direction. We take the median valid depth in the central third of the image along both dimensions and clip it to a bounded range, obtaining $\bar{d}_s$. A nominal distance is used when valid depth is unavailable.

To re-examine the supporting scene, we generate local viewpoint proposals in the forward, backward, left, and right directions relative to the original camera heading. Let Δ_s denote the corresponding displacement vectors, directed along $\pm\mathbf{g}_s$ and $\pm\mathbf{g}_s^{\perp}$, where $\mathbf{g}_s^{\perp} = (-g_{s,2}, g_{s,1})^{\top}$. The coarse scene anchor and proposed camera positions are

$$\mathbf{a}_s = \mathbf{p}_s + \bar{d}_s \mathbf{g}_s, \qquad \mathcal{P}_s^{\mathrm{prop}} = \{\mathbf{p}_s + \boldsymbol{\delta} : \boldsymbol{\delta} \in \Delta_s\}, \tag{9}$$

where the anchor approximates the observed scene region. Changing the camera position allows the agent to inspect this region from different distances and lateral perspectives, potentially revealing details that are obscured or ambiguous in the original observation.

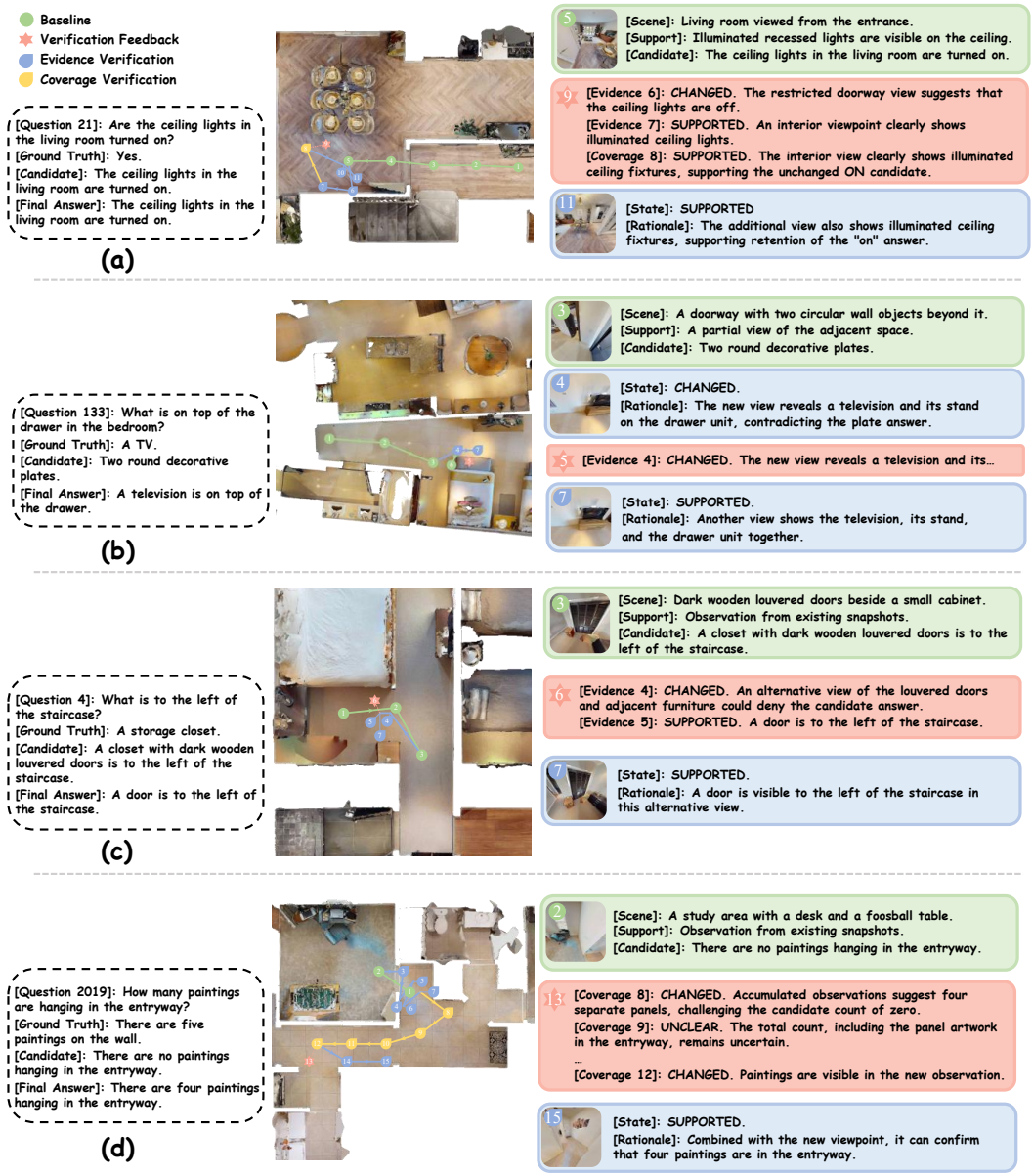


Figure 13: Extended qualitative analysis of Sure-EQA. (a) Correct to correct: retaining a correct answer despite conflicting verification judgments. (b) Incorrect to correct: correcting object identification with additional evidence. (c) Correct to incorrect: losing semantic specificity during revision. (d) Incorrect to incorrect: correcting an absence claim while retaining a counting error.

Let Π_t snap a proposal to the nearest free cell in the current navigation map, returning $\perp$ if no such point is available. After snapping, we retain positions that are sufficiently separated from the original camera:

$$\mathcal{P}_{s,t}^{E} = \{\mathbf{p} \in \{\Pi_t(\mathbf{p}') : \mathbf{p}' \in \mathcal{P}_s^{\text{prop}}\} \setminus \{\perp\} : \|\mathbf{p} - \mathbf{p}_s\|_2 \geq \delta_E\}, \tag{10}$$

where δ_E prevents snapping from collapsing a proposal back to the original viewpoint. Each retained position is assigned a unit viewing direction toward the same scene anchor:

$$\mathbf{g}_s(\mathbf{p}) = \frac{\mathbf{a}_s - \mathbf{p}}{\|\mathbf{a}_s - \mathbf{p}\|_2}, \qquad \mathbf{p} \in \mathcal{P}_{s,t}^{E}. \tag{11}$$

This keeps each candidate view directed toward the supporting scene region. If none of the proposed viewpoints survives this filtering, the underlying observation planner searches for another feasible viewpoint around the same anchor.

Coverage viewpoint construction. Coverage viewpoints are constructed from reachable frontiers of a truncated signed distance function (TSDF) map (Ren et al., 2024; Yang et al., 2025b). The map is incrementally updated using depth observations and camera poses. Within the navigable slice, free-space regions near the executed trajectory are marked as explored, and frontiers are extracted along the boundaries between explored and unexplored regions. DBSCAN groups spatially adjacent frontier pixels into candidate regions (Ester et al., 1996). Small clusters are discarded, while regions with an excessively large angular extent are divided into smaller sub-frontiers. Let the resulting geometrically generated frontier set be

$$\mathcal{F}_t = \left\{(r_j, \mathbf{b}_j, \mathbf{g}_j^F, I_j^F)\right\}_{j=1}^{m_t}, \tag{12}$$

where r_j is a frontier region, $\mathbf{b}_j$ is its navigable boundary position, $\mathbf{g}_j^F$ is the viewing direction toward the associated unexplored region, and I_j^F is an already acquired frontier-facing image. These images depict the visible boundaries or entrances available at the current exploration state. Every frontier in $\mathcal{F}_t$ generates a coverage viewpoint:

$$\mathcal{U}_t^{C,\text{geom}} = \left\{u_j^C = (\mathbf{b}_j, \mathbf{g}_j^F) \mid j = 1, \ldots, m_t\right\}. \tag{13}$$

Each coverage viewpoint directs the agent toward the corresponding boundary position to acquire an observation of the associated region. These geometrically generated views provide the coverage candidates for subsequent verification. The frontier set is refreshed as exploration updates the map.

Frontier construction is geometric and does not apply question- or candidate-conditioned semantic filtering. The acquired observations are evaluated with respect to the candidate during verification. Equation 7 selects the next executable viewpoint from the two verification types. After an observation is acquired, the corresponding verifier evaluates it through Equation 3.

A.3.2 Geometric Weighting and Evidence Aggregation

Checks acquired at camera poses similar to a target u provide more relevant findings for assessing that target. Each check's contribution is also weighted by its geometric novelty relative to earlier checks of the same type and outcome. This weighting limits redundant contributions from repeated checks at nearly identical poses. This section specifies the kernel and weights used in Equation 5, together with the novelty term used for view selection in Equation 7.

For a view u, let (p_x, p_y) be its horizontal camera position and (g_x, g_y) its unit viewing direction. We define

$$\phi(u) = [p_x/L, p_y/L, g_x, g_y]^\top, \qquad k(u, u') = \exp\left(-\frac{\|\phi(u) - \phi(u')\|_2^2}{2\sigma^2}\right), \tag{14}$$

where $L > 0$ sets the spatial scale and $\sigma > 0$ controls the decay of similarity. Proposed views use their planned poses, whereas completed checks use the poses at which observations were actually acquired.

For a set $\mathcal{J}$ of completed-check indices, define novelty as

$$\nu(u; \mathcal{J}) = 1 - \max_{j \in \mathcal{J}} k(u, u_j), \tag{15}$$

with the maximum over an empty history defined as zero. Thus, the first check has novelty one, while a view identical to a recorded view has novelty zero.

When check i yields `SUPPORTED` or `CHANGED`, its evidence weight compares it with earlier checks of the same verification type and the same outcome:

$$\mathcal{J}_i = \{j < i : d_j = d_i,\ z_j = z_i\}, \qquad w_i = \log(1 + \nu(u_i; \mathcal{J}_i))\,. \tag{16}$$

The history contains only checks of the current verification round. Each weight is computed before recording the new check and remains fixed thereafter. Repeated identical outcomes at the same pose therefore add zero weight, while an earlier opposite outcome does not discount the first occurrence of the new outcome at that pose. These fixed weights enter the two evidence quantities in Equation 5, which determine the verification risk score in Equation 4. Only checks with acquired observations contribute to these quantities.

View selection uses a different history. Let $\mathcal{I}_d^{\text{all}}$ index all completed checks of type d in the current verification round, including `UNCLEAR`. For a target of type d, the novelty in Equation 7 is

$$\nu(u) = \nu(u; \mathcal{I}_d^{\text{all}}). \tag{17}$$

An inconclusive check therefore contributes no supporting or answer-changing evidence, but still reduces the priority of revisiting a similar view. The floor ϵ applies only to view selection; it does not assign additional weight to repeated evidence.

A.3.3 Target Updates and Verification Control

The evidence and coverage branches use the same acceptance and round-termination rules. For each type $d \in \{E, C\}$, $\mathcal{U}^d$ contains pending targets for further observation, and $\mathcal{T}^d$ contains all targets in the current verification round, including those already checked. As defined in Appendix A.3.2, $\mathcal{I}_d^{\text{all}}$ indexes all completed checks of type d in the current round, including `UNCLEAR` outcomes. Thus, $\mathcal{T}^d = \mathcal{U}^d \cup \{u_i : i \in \mathcal{I}_d^{\text{all}}\}$. The currently reachable targets in $\mathcal{U}^E \cup \mathcal{U}^C$ form the selection set $\mathcal{U}$ used in Section 3.3. Each target retains its verification type, which determines its risk score and the verifier applied to its observations.

After a target is checked, it is removed from $\mathcal{U}^d$ and retained in $\mathcal{T}^d$ for the rest of the round. Its acquired viewpoint u_i, observation, outcome, and rationale remain in the verification history. A completed target is evaluated at its acquired pose, following Appendix A.3.2. Newly generated reachable frontier targets are added to both $\mathcal{U}^C$ and $\mathcal{T}^C$ as observations update the geometric map.

After incorporating each new verification result, we update the scores of all targets in $\mathcal{T}^d$ using Equations 4 and 5. Each $r_d(u)$ therefore reflects the verification history accumulated so far in the current round, including results obtained after u was checked. Both branches use

$$R_d = \begin{cases} \max\limits_{u \in \mathcal{T}^d} r_d(u), & \mathcal{T}^d \neq \varnothing, \\ 0, & \mathcal{T}^d = \varnothing, \end{cases} \qquad d \in \{E, C\}. \tag{18}$$

A branch has zero risk when it has no verification targets. Exhausting its pending set leaves its completed targets in the risk evaluation. A previously checked target therefore continues to prevent acceptance while its current risk exceeds the threshold.

The controller evaluates acceptance using these updated branch risks. It accepts the candidate when $R_E \leq \tau_E$ and $R_C \leq \tau_C$. These two inequalities are equivalent to the acceptance criterion in Section 3.3, which requires every target of both types to satisfy its corresponding risk threshold. If either branch has no pending targets and its risk remains above its threshold, the round ends immediately with verification feedback, even when the other branch still has pending targets. Otherwise, verification continues while the interaction budget permits. A `CHANGED` outcome contributes to the risk calculation without independently triggering another observation or candidate replacement.

The candidate $\hat{y}$ and its original support $\mathcal{S}$ remain fixed throughout the round. The buffer retains all acquired images $\mathcal{O}_{\text{ver}}$ and their verification record $\mathcal{H}_{\text{ver}}$. For each executed image-level check i, the record contains the observation ID, verification type d_i, actual camera pose u_i, judgment z_i, and explanation e_i. The observation ID links each record to its acquired image in $\mathcal{O}_{\text{ver}}$. When a round ends without acceptance, this buffer is committed through Equation 8. The update includes all three

outcome types and only observations that were actually acquired. The answerer then reconsiders the question using the updated memories. It may propose a candidate and supporting observations or request further exploration. A new verification round begins only when a candidate and its support are available and interaction budget remains. It starts with fresh target sets, empty verification histories, and a fresh buffer. Previously committed observations and structural records remain available in memory.

Exploration and verification share the same episode budget. If the budget is exhausted before verified acceptance, the observations and records collected so far are also committed through Equation 8, including those from an incomplete round. The answerer uses the updated memories to generate the final answer, which is returned as budget-triggered answering without another verification round.

A.4 Implementation Details

A.4.1 VLM Deployment

Sure-EQA uses open-source VLMs through vLLM (Kwon et al., 2023). To reduce repeated processing during exploration and verification, we share the model service across roles and reuse image preparation results.

- **Persistent shared service.** The underlying agent's VLM-based exploration planner selects where to seek task-relevant observations. It shares a resident model service with the answerer, evidence verifier, and coverage verifier, using role-specific prompts. This avoids deploying a separate model instance for each role. The navigation controller executes movement toward the selected targets.
- **Image encoding reuse.** The same observation may be used repeatedly by the memory managers, answerer, and verifiers. When an observation is first prepared for a VLM request, the client compresses the image into PNG format and converts it into a Base64 string for transmission. This string is stored in a local cache. When the same retained image is needed again, the client directly retrieves the cached string instead of repeating compression and conversion. The cache removes least recently used entries when its capacity is reached, keeping memory consumption bounded. This reuse reduces CPU-side request preparation overhead.
- **Compact structured outputs.** Prompts request concise rationales followed by the required structured output. Memory summaries and most selection or answering requests use generation limits of 192–256 tokens, while high-level planning and verification judgments use limits of 512 tokens.

Table 6: Runtime and estimated model computation of Baseline exploration and Sure-EQA verification using Qwen3-VL 8B on EXPRESS-Bench. VLM/LLM denotes image-conditioned/text-only requests. Average steps refer to the corresponding phase per episode.

Method	VLM/LLM Calls/Step	Step latency (s)	Avg. Steps per Episode	TFLOPs /Step
Exploration (Baseline)	3.62/5.05	81.6	6.95	907
Verification (Ours)	3.24/2.46	38.1	3.57	185

Table 6 compares Baseline exploration and Sure-EQA verification on the same hardware. Baseline exploration involves repeated structural-memory construction and lengthy text generation, while verification focuses on assessing the current candidate using newly acquired observations. Verification averages 3.24 image-conditioned and 2.46 text-only requests per step, compared with 3.62 and 5.05, respectively, for baseline exploration. We estimate FLOPs from logged image sizes, token counts, and cache usage, covering visual encoding and language-model inference only. For each phase, FLOPs per step are calculated by dividing the total estimated FLOPs by the total number of executed steps across episodes. Verification requires 38.1 seconds and 185 TFLOPs per step, compared with 81.6 seconds and 907 TFLOPs for baseline exploration, corresponding to reductions of 53.3% and 79.6%, respectively. With an average of 3.57 verification steps per episode, Sure-EQA improves C^* by 7.65 points and C by 4.73 points over Baseline on EXPRESS-Bench. These results show that

Table 7: Verification parameters used on both benchmarks.

Parameter	Meaning	Value
α_0	Answer-change smoothing mass	2
β_0	Supporting smoothing mass	1
τ_E, τ_C	Verification thresholds	0.53, 0.53
L	Spatial scale in the view descriptor	2.0 m
σ	Kernel bandwidth	0.75
ϵ	Novelty floor for view selection	0.05

substantial answer-quality gains can be obtained through a small number of targeted verification steps, each with substantially lower latency and model computation than a baseline exploration step.

A.4.2 Configuration Setup

The following settings specify observation acquisition, geometric mapping, navigation, and the shared interaction budget. Verification parameters are listed separately to connect the implementation with the scoring and selection rules in the main text.

Observation setup. The camera is positioned $1.5\,\mathrm{m}$ above the floor, with a horizontal field of view of $120°$. RGB and depth observations are rendered at 1280×1280 resolution, and RGB inputs are resized to 640×640 for VLM inference. Initially, seven views are captured with heading offsets from $-120°$ to $120°$ in $40°$ increments, providing full surrounding coverage through their combined fields of view. Subsequent steps collect three views at offsets of $-60°$, $0°$, and $60°$. Camera height remains fixed throughout exploration and verification.

Mapping and navigation. The TSDF map uses a voxel size of $0.1\,\mathrm{m}$, and the explored-region distance parameter is set to $1.7\,\mathrm{m}$. Frontier proposals are extracted along the explored-space boundary. The navigation controller executes translations of up to $1.0\,\mathrm{m}$ per step. Each verification step acquires three images at the same position with different viewing directions. The images are evaluated separately, each producing its own verifier outcome and rationale. Each image-level judgment is recorded as a separate check at its actual camera pose. Thus, one verification step comprises three verifier checks, and the check index i in the aggregation equations enumerates image-level judgments rather than observation steps. The reported verification-step counts refer to observation steps, with each three-image acquisition counted once. For the main benchmark evaluations, exploration and verification share a budget of 50 interaction steps per episode. When the shared episode budget is exhausted before verified stopping, we apply the feedback procedure in Appendix A.3.3 and return the regenerated answer without further navigation or verification. This outcome is recorded as budget-triggered answering.

Verification parameters. Table 7 lists the parameters used for verification. We use the same values on both benchmarks. The descriptor and kernel are defined in Appendix A.3.2. We set $0 < \tau_E, \tau_C < r_0$, where $r_0 = \alpha_0/(\alpha_0 + \beta_0)$ is the initial target score. A generated target therefore requires supporting verification results before its risk can fall to the acceptance threshold. We use the training data of OpenEQA and EXPRESS-Bench as development sets for selecting verification thresholds and other method hyperparameters. The selected configuration is fixed before evaluation and used across all main comparisons. For each control rule in the threshold sensitivity analysis, we vary the shared branch threshold while keeping all other method hyperparameters fixed at their development-set values.

Ablation settings. For the verification selection and feedback ablations in Table 3, *w/o Risk Score* removes the risk term from viewpoint selection while retaining the joint acceptance criterion. *w/o Structural Record* omits the addition of verification records to the structural memory used for subsequent reasoning. *w/o Visual Feedback* omits the addition of newly acquired verification images to visual memory, while the verifiers still receive these images when evaluating the candidate.

A.4.3 CANDIDATE GENERATOR

Sure-EQA operates on the candidate answer, supporting observations, and memory supplied by an underlying exploration agent. Our primary implementation uses Pred-EQA (Yuan et al., 2026) in this role, which we refer to as the candidate generator. Its exploration and memory modules provide task-relevant visual observations and structural information about the explored environment. From these memories, the VLM-based answerer produces a candidate answer and identifies its supporting observations when the agent proposes to stop.

A stop proposal transfers viewpoint selection to the verification-based policy. The candidate and its original support remain fixed while Sure-EQA selects viewpoints and evaluates new observations against the proposed interpretation. The navigation controller executes the selected movements. When a round ends without acceptance, the acquired images and verification records are committed through Equation 8. If interaction budget remains, the answerer uses the updated memories to propose another candidate or request further exploration. Round transitions and budget handling follow Appendix A.3.3.

A.5 BENCHMARK DETAILS

We evaluate Sure-EQA on two active EQA benchmarks, A-EQA and EXPRESS-Bench. Both use indoor environments from HM3D (Ramakrishnan et al., 2021) and require agents to acquire visual evidence through exploration.

- **A-EQA.** The active setting of OpenEQA (Majumdar et al., 2024) comprises 557 questions across 63 HM3D scenes. Following prior work (Yuan et al., 2026), we use the publicly released 184-question evaluation subset. The questions are open-ended and cover seven categories: object recognition, attribute recognition, object state recognition, object localization, spatial understanding, functional reasoning, and world knowledge. Answering these questions requires identifying relevant scene content and integrating observations acquired from different viewpoints.
- **EXPRESS-Bench.** EXPRESS-Bench (Jiang et al., 2025) contains 777 exploration trajectories and 2,044 question–trajectory pairs. For consistency with prior evaluations, we reuse the 409-question evaluation subset used by the compared methods. Its questions span seven categories: state, knowledge, location, attribute, counting, existence, and object. In addition to answer correctness, the benchmark evaluates whether the acquired observations support the generated answer. Its reference trajectories and target locations also support the assessment of exploration efficiency and terminal navigation accuracy.

In the main benchmark evaluations, each agent receives a question and starts from a designated pose. It navigates the environment, collects observations, and decides when to return an answer. Reference answers, annotated trajectories, and target locations are reserved for evaluation and are not provided to the agent for planning or verification. The manual evidence insertion in the pilot study is a separate diagnostic intervention.

A.6 METRICS DETAILS

We reuse Pred-EQA's (Yuan et al., 2026) scoring method, following the OpenEQA answer-matching protocol for A-EQA and the EXPRESS-Bench evaluation protocol for EXPRESS-Bench. Let N denote the number of evaluated episodes and p_i the total traveled path length in episode i. For Sure-EQA, this length includes both answer-seeking exploration and verification movements. All answer-quality and path-weighted scores are reported on a 0–100 scale, while terminal distance is reported in meters.

A-EQA. We report LLM-Match and LLM-SPL following OpenEQA and prior active-EQA evaluation (Majumdar et al., 2024; Yuan et al., 2026). A GPT-4 evaluator compares the predicted answer with the reference answer and assigns a semantic correctness score $s_i \in \{1, 2, 3, 4, 5\}$ based on the OpenEQA prompt. LLM-Match linearly maps this score to the interval $[0, 1]$ before averaging:

$$\text{LLM-Match} = \frac{100}{N} \sum_{i=1}^{N} \frac{s_i - 1}{4}. \tag{19}$$

This semantic evaluation allows equivalent answers to receive credit without requiring an exact textual match. LLM-SPL additionally weights answer correctness by path efficiency:

$$\text{LLM-SPL} = \frac{100}{N}\sum_{i=1}^{N}\frac{s_i - 1}{4}\frac{l_i}{\max(p_i, l_i)}. \tag{20}$$

Here, l_i denotes the minimum length along the reference trajectory sufficient to acquire the evidence needed to answer the question. The path factor equals one when the agent travels no farther than this reference length and decreases for longer trajectories. Thus, additional verification contributes positively only when its improvement in answer correctness offsets its navigation cost.

EXPRESS-Bench. We report C, C^*, E_{path}, and d_T following the EXPRESS-Bench protocol (Jiang et al., 2025). The evaluation separates answer correctness from visual grounding. Each episode receives an answer correctness score $s_i \in \{1, 2, 3, 4, 5\}$ and a grounding score $\delta_i \in \{0, 0.5, 1\}$. A grounding score of one indicates relevant observations with an accurate scene description. A score of 0.5 indicates relevant observations but an inaccurate description, whereas zero indicates irrelevant observations or hallucinated scene content. These grounding judgments are distinct from semantic agreement with the reference answer.

The grounded correctness score combines both components:

$$C = \frac{100}{N}\sum_{i=1}^{N}\frac{s_i\delta_i}{5}. \tag{21}$$

Consequently, an answer that matches the reference but lacks relevant visual support cannot receive full credit. To evaluate answer correctness independently of grounding, C^* omits the grounding factor:

$$C^* = \frac{100}{N}\sum_{i=1}^{N}\frac{s_i}{5}. \tag{22}$$

This is equivalent to setting $\delta_i = 1$ for every episode. Unlike the A-EQA normalization, EXPRESS-Bench divides the raw correctness score by five without subtracting one.

The path-aware score further incorporates navigation efficiency:

$$E_{\text{path}} = \frac{100}{N}\sum_{i=1}^{N}\frac{s_i\delta_i}{5}\frac{l_i}{\max(p_i, l_i)}. \tag{23}$$

For EXPRESS-Bench, l_i denotes the geodesic distance from the initial position to the reference target. Importantly, E_{path} weights grounded correctness C, rather than answer-only correctness C^*. It therefore rewards answers that are accurate, visually supported, and obtained through efficient exploration.

Finally, terminal navigation accuracy is measured by

$$d_T = \frac{1}{N}\sum_{i=1}^{N}\text{dist}_{\text{geo}}\left(\mathbf{p}_i^{\text{end}}, \mathbf{p}_i^{\text{goal}}\right), \tag{24}$$

where $\mathbf{p}_i^{\text{end}}$ is the agent's final position, $\mathbf{p}_i^{\text{goal}}$ is the reference target position, and dist_{geo} denotes geodesic distance. Lower d_T indicates closer proximity to the target, but does not by itself guarantee a correct or grounded answer. We therefore interpret it alongside the other three metrics.

A.7 Agent Prompt Details

We provide the prompt templates for candidate answering, evidence verification, and coverage verification, together with the structured feedback context used for subsequent answering. Each template specifies its instructions, input fields, and output format.

Figure 14 presents the candidate answerer prompt, which supports both initial candidate generation and reconsideration after verification feedback. When evidence is insufficient and budget remains,

the answerer can request further exploration. Verification begins only after a candidate answer and its supporting observation IDs are available. At budget exhaustion, the answerer returns a final answer using the available memory.

Figures 15 and 16 present the two verifier prompts. The evidence verifier re-examines the candidate's visual support, while the coverage verifier checks newly observed regions for information that could revise it. Both receive the fixed candidate, original supporting images, a new observation, and structural memory, and return a verification outcome z_i and a visual rationale e_i. Risk aggregation, view selection, and stopping decisions are handled by the controller.

Figure 17 shows the verification feedback context assembled by the controller from the execution log. The context contains the verification record $\mathcal{H}_{\text{ver}}$, together with the candidate under examination, its original supporting observation IDs, and the reason the round ended. Each check record contains (d_i, u_i, z_i, e_i) and an observation ID linking it to an acquired image in $\mathcal{O}_{\text{ver}}$. The images are committed to visual memory, while the verification records are retained in structural memory. This context allows the answerer to reconsider its candidate using the acquired images and the findings associated with them. When the budget is exhausted, the context also retains the available observations and records from an incomplete round.

Task: You are the candidate answerer of an embodied question answering agent. Use the question, visual observations, and structural memory to propose an answer and identify the observations supporting it. Your answer is a candidate for subsequent verification, not a verified stopping decision.

Instructions:
1. Identify what the question asks: the relevant object or region, its identity, location, attributes, state, count, or relationships. Match the specificity of your answer to the question.
2. Inspect all supplied observation images and their IDs. Use structural memory to interpret the scene and previous exploration, but do not treat an earlier prediction or room-level association as a directly observed fact.
3. Combine compatible visual clues when needed. Check that observations concern the intended object and region. Do not fill cropped, occluded, or unreadable details with plausible guesses.
4. If verification feedback is present, reconsider the previous candidate using the newly committed observations and records. Distinguish evidence calling for a change from an inconclusive check; an unaccepted stop proposal does not by itself prove that the answer is false.
5. When a candidate can be proposed, return a concise answer and valid supporting observation IDs. If the evidence is insufficient and exploration budget remains, return Continue Exploration and state what is missing. This rule applies both during ordinary exploration and after verification feedback. Do not invent observation IDs.
6. Continue Exploration returns control to the Planner for further exploration; it does not start a verification round. A new round begins only after a candidate answer and its supporting observation IDs are returned, provided budget remains.

Target Question: [QUESTION]

Structural Memory:
[EXPLORED REGIONS, OBSERVED FACTS, CURRENT PLAN, AND UNRESOLVED ITEMS]

Visual Memory:
Observation [ID_1]: [IMAGE_1]
Observation [ID_2]: [IMAGE_2]
[ADDITIONAL AVAILABLE OBSERVATIONS]

Verification Feedback:
[COMMITTED FEEDBACK CONTEXT, IF AVAILABLE; OTHERWISE NONE]

Output Format:
Rationale: [BRIEF DESCRIPTION OF THE VISIBLE SUPPORT OR MISSING EVIDENCE]
If a candidate can be proposed, or the episode budget is exhausted:
Answer: [CONCISE CANDIDATE ANSWER]
Supporting Observation IDs: [ID_1, ID_2, ...]
If evidence is insufficient and exploration budget remains (including after verification feedback):
Continue Exploration

Budget-Triggered Answering: If the caller explicitly states that the episode budget is exhausted, return the best answer supported by the supplied memory and acknowledge unresolved details. Do not request another action or claim that verification has passed.

Figure 14: Candidate Answerer prompt for initial answering and reconsideration after verification feedback.

Task: You are the evidence verifier. Re-examine the visual basis of a fixed candidate answer using its original supporting images and one newly acquired observation from an alternative viewpoint. Determine whether this check supports retaining the candidate, calls for a revision, or remains inconclusive.

Instructions:
1. Read the question and identify the specific scene claim made by the candidate. Inspect the original supporting images rather than assuming that their selection guarantees correctness.
2. Compare the new observation with the original support. Establish whether it depicts the same relevant object or region, and inspect the identity, location, attributes, state, and relationships needed by the question.
3. Use structural memory for context, but base the outcome on visible evidence. A familiar room layout or common object association is not sufficient to resolve an ambiguous visual claim.
4. Treat cropping, occlusion, unreadable content, or an uncertain correspondence between views as inconclusive when they prevent a decision. An object falling outside the new field of view does not by itself show that the candidate is wrong.
5. Keep the candidate and original supporting set fixed. Report the outcome of this check and a short visual rationale; do not replace the answer, choose a navigation action, or decide whether the overall verification round should stop.

Outcome Definitions:
SUPPORTED: The compared views provide sufficient visual support for retaining the candidate on the claim being checked.
CHANGED: Visible evidence contradicts, corrects, or materially completes the candidate and therefore calls for a revision.
UNCLEAR: The available views do not resolve the relevant claim, including ambiguity, occlusion, or uncertain object or region correspondence.

Target Question: [QUESTION]

Frozen Candidate: [CANDIDATE ANSWER]

Original Supporting Observations:
Observation [SUPPORT_ID_1]: [SUPPORT_IMAGE_1]
[ADDITIONAL ORIGINAL SUPPORT IDS AND IMAGES]

New Verification Observation:
Observation [NEW_ID]: [NEW_IMAGE]

Structural Memory: [EXISTING STRUCTURAL MEMORY]

Output Format:
State: [SUPPORTED, CHANGED, OR UNCLEAR; SELECT EXACTLY ONE]
Rationale: [ONE OR TWO SENTENCES IDENTIFYING THE OBSERVATION IDS AND THE VISIBLE FACTS OR UNRESOLVED DETAILS THAT JUSTIFY THE STATE]

Figure 15: Evidence Verifier prompt for re-examining the visual support of a candidate answer.

Task: You are the coverage verifier. Assess whether information from the newly observed region calls for revising the fixed candidate answer. Base the assessment on the acquired observation and visually grounded inferences about the surrounding region.

Instructions:
1. Identify which new information could change the answer to the question, such as another relevant object, a different location, an additional counted instance, or a conflicting attribute or state.
2. Inspect the new observation together with the original supporting images. Determine whether any newly visible fact changes the candidate's meaning or completeness, rather than merely adding unrelated scene detail.
3. Use structural memory to interpret the inspected region. Do not assume that an unexplored frontier contains a target, or treat an expected room function as observed evidence.
4. Limit negative conclusions to the visible, inspected content. Not seeing a relevant object in this view does not prove its absence from the whole scene. If a potentially decisive part of the region cannot be assessed, return UNCLEAR.
5. Keep the candidate fixed and report one outcome with a concise visual rationale. Do not declare the entire environment covered, select another frontier, update the answer, or determine the global stopping decision.

Outcome Definitions:
SUPPORTED: The observed content and visually grounded inferences about the surrounding region provide sufficient grounds to retain the candidate.
CHANGED: The new region reveals information that contradicts, corrects, or materially completes the candidate answer.
UNCLEAR: The current observation does not provide sufficient visual context to assess whether the relevant region or its extension could contain information requiring a revision of the candidate.

Target Question: [QUESTION]

Frozen Candidate: [CANDIDATE ANSWER]

Original Supporting Observations:
Observation [SUPPORT_ID_1]: [SUPPORT_IMAGE_1]
[ADDITIONAL ORIGINAL SUPPORT IDS AND IMAGES]

New Region Observation:
Observation [NEW_ID]: [NEW_IMAGE]

Structural Memory: [EXISTING STRUCTURAL MEMORY]

Output Format:
State: [SUPPORTED, CHANGED, OR UNCLEAR; SELECT EXACTLY ONE]
Rationale: [ONE OR TWO SENTENCES IDENTIFYING THE OBSERVATION IDS AND EXPLAINING THE ANSWER-CHANGING FACT, THE RESOLVED LOCAL CHECK, OR THE REMAINING VISUAL AMBIGUITY]

Figure 16: Coverage Verifier prompt for checking newly observed regions for information that could revise the answer.

Context Type: Controller-assembled verification feedback. This is a structured context block supplied to the next Candidate Answerer, not a separate prompt that invokes another model.

Purpose: Preserve the candidate under examination, the observations acquired during verification, and the outcomes explaining why the stop proposal was not accepted. The next answer is generated from the updated visual and structural memories.

Assembly Rules:
1. Retain the candidate answer and its original supporting observation IDs. Record that the stop proposal was not accepted; this does not automatically establish that every claim in the candidate is false.
2. Include each executed check with its verification type, acquired observation ID, recorded camera pose, outcome, and visual rationale. Retain SUPPORTED, CHANGED, and UNCLEAR records; exclude proposals that were never executed.
3. Take observation IDs and poses from the execution log rather than asking a model to invent them. Link each record to its actual acquired image. Store these images in visual memory and the records in structural memory.
4. If the episode budget ends during an incomplete round, preserve the available records and identify the return as budget-triggered answering, not a successful verification result.

Target Question: [QUESTION]

Candidate Under Examination: [CANDIDATE ANSWER]

Original Supporting Observation IDs: [SUPPORT_ID_1, SUPPORT_ID_2, ...]

Round Summary:
Stop proposal: NOT ACCEPTED
End reason: [CONTROLLER-RECORDED REASON; INDICATE BUDGET EXHAUSTION WHEN APPLICABLE]

Records:
Check [i]: EVIDENCE; Observation [ID_i]; Position ([x_i], [y_i]); Heading ([gx_i], [gy_i]); State [OUTCOME_i]; Rationale [VISUAL RATIONALE_i].
Check [j]: COVERAGE; Observation [ID_j]; Position ([x_j], [y_j]); Heading ([gx_j], [gy_j]); State [OUTCOME_j]; Rationale [VISUAL RATIONALE_j].
[ONE RECORD PER EXECUTED CHECK, IN EXECUTION ORDER]

Committed Visual Observations:
Observation [ID_i]: [ACQUIRED IMAGE_i]
Observation [ID_j]: [ACQUIRED IMAGE_j]
[ALL OTHER ACQUIRED VERIFICATION OBSERVATIONS FROM THIS ROUND]

Use in Subsequent Answering: Reassess the previous candidate against the updated visual memory. Use CHANGED records to examine possible corrections and UNCLEAR records to identify unresolved details. Treat the records as prior judgments to check against the images, not as new visual facts. Follow the Candidate Answerer's output and budget rules: return a candidate and its supporting observation IDs, or Continue Exploration if evidence remains insufficient and budget is available. Do not mark an answer as verified.

Figure 17: Verification feedback context assembled by the controller and supplied to subsequent answering.